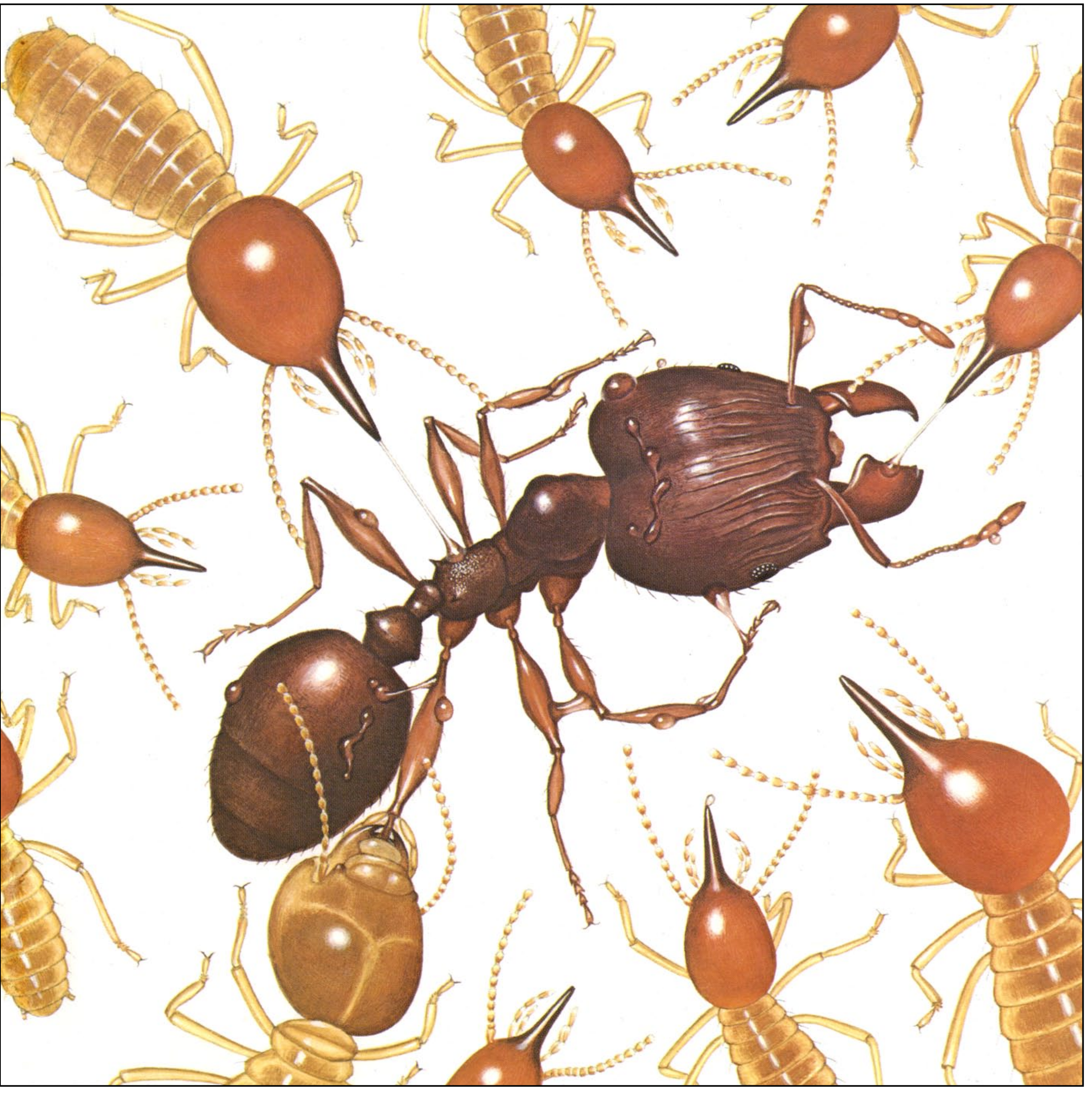
INVESTIGACION Y CIENCIA

Edición en español de **SCIENTIFIC AMERICAN**



ARMAS QUIMICAS DE LOS TERMES

Octubre 1983

350 Ptas.

Los espacios en gris
corresponden a publicidad
en la edición impresa

Número 85
Octubre 1983

INVESTIGACION Y
CIENCIA
Edición en español de SCIENTIFIC AMERICAN

EDITA Prensa Científica, S. A.
Calabria, 235-239
Barcelona-29 (ESPAÑA)

Colaboradores de este número:

Asesoramiento y traducción:

Isabel García Acha y J. Rodríguez Villanueva: *Purificación y fabricación de interferones humanos*; Manuel Puigcerver: *Campos magnéticos en el cosmos* y *Materia interestelar en meteoritos*; Joandomènec Ros: *Las defensas químicas de los termes*; Luis Bou y Salvador Barberá: *Decisiones racionales colectivas*; Teodoro Falcón Márquez: *Iglesias noruegas de pies derechos*; Amando García Rodríguez: *Tipografía digital*; Luis Bou: *Juegos matemáticos*; J. Vilardell: *Taller y laboratorio.*

Libros:

Luis García Ballester



LA PORTADA

La ilustración de la portada muestra un grupo de termes que intenta inmovilizar una hormiga invasora de la especie *Pheidole megacephala.* Un termes obrero sujeta al intruso por su pata posterior derecha, pero los principales defensores del termitero son miembros de la casta de los termes soldados. Los soldados de esta especie de termes, *Trinervitermes bettonianus*, son de dos tamaños y defienden el nido mediante una secreción irritante y viscosa (véase "Las defensas químicas de los termes", por Glenn D. Prestwich, en este mismo número). El intruso ha sido alcanzado por varias rociadas viscosas. La mayoría de los pequeños soldados han agotado ya su reserva de visco, pero la defensa de los soldados ha inmovilizado dos de las patas de la hormiga.

Suscripciones:
Prensa Científica, S. A.
Calabria, 235-239
Barcelona-29 (España)
Teléfono 322 05 51 ext. 37

Condiciones de suscripción:
España:
Un año (12 números): 3.850 pesetas
Extranjero:
Un año (12 números): 52 U.S. $
Ejemplar atrasado ordinario:
350 pesetas
Ejemplar atrasado extraordinario:
450 pesetas

Distribución para España
Distribuciones de Enlace, S. A.
Ausias March, 49 - Barcelona-10

Distribución para los restantes países:
Editorial Labor, S. A.
Calabria, 235-239 - Barcelona-29

Publicidad:
Madrid:
Gustavo Martínez Ovin
Avda. de Moratalaz, 137 - Madrid-30
Tel. 430 84 81
Cataluña:
Lourdes Padrós
Manuel Girona, 61, 2.° - Barcelona-34
Teléfono 204 45 83

Controlado por O.J.D.
OJD DIFUSION PUBLICIDAD

PROCEDENCIA DE LAS ILUSTRACIONES

Pintura de la portada de Tom Prentiss

Página	Fuente
8-16	Juan Cotoner y Josep Font
18-19	Hoffman-La Roche, Inc.
20-25	Bunji Tagawa
29	High Altitude Observatory de la Universidad de Colorado en Boulder
30-39	George V. Kelvin
43	Glenn MacPherson, Universidad de Chicago
44-49	Walken Graphics
50	Mitsuo Ohtsuki, Universidad de Chicago
51-53	Walken Graphics
55	Barbara L. Thorne, Museo de Zoología Comparativa, Universidad de Harvard
56	Glenn D. Prestwich, Universidad estatal de Nueva York en Stony Brook
57-59	Tom Prentiss
60	Barbara L. Thorne, Museo de Zoología Comparativa, Universidad de Harvard
61	Manfred Kaib, Universidad de Bayreuth
62	Barbara L. Thorne, Museo de Zoología Comparativa, Universidad de Harvard (*arriba*);

Página	Fuente
62	Glenn D. Prestwich, Universidad estatal de Nueva York en Stony Brook (*abajo*)
65-72	Ilil Arbel
75	Ronald L. Sack, Universidad de Idaho
76	Mittet Foto A/S
77-79	J. Dyck Fledderus
80-81	Ronald L. Sack, Universidad de Idaho
82-83	J. Dyck Fledderus
85	Autologic, Inc.
86-87	Kris Holmes y Edward Bell
88	Allen Beechel (*arriba*); Charles Bigelow, Universidad de Stanford (*abajo*)
89-92	Allen Beechel
93	John E. Warnock, Xerox Corporation
94-96	Allen Beechel
98	Peter Karow, URW Unternehmensberatung, Hamburgo, y Donald E. Knuth, Universidad de Stanford, cortesía de *Visible Language*
100-104	Ilil Arbel
107-112	Michael Goodman

ISSN 0210-136X
Dep. legal: B. 38.999-76
Fotocomposición Tecfa
Pedro IV, 160 - Barcelona-5
Fotocromos reproducidos por GINSA, S.A.
Imprime GRAFESA
Gráfica Elzeviriana, S. A.
Napoles, 249 - Tel. 207 40 11
Barcelona-13
Printed in Spain - Impreso en España

Los autores

JOSEP FONT ("Química en fase sólida") es catedrático de química orgánica de la Facultad de Ciencias de la Universidad Autónoma de Barcelona (UAB) y decano de la misma. Cursó la carrera de química en la Universidad de Barcelona, por la que se doctoró en 1965, bajo la dirección del extinto profesor José Pascual, con quien colaboró posteriormente en síntesis de alcoholes terpénicos monocíclicos. Entre 1967 y 1969 trabajó con O. P. Strausz, en Edmonton, Canadá, en la determinación de intermedios químicos generados por vía fotoquímica. De regreso a España, trabajó con F. Serratosa en la síntesis de diversas moléculas orgánicas, y en polímeros funcionalizados con J. Castells, que fue el que introdujo la idea de utilizar los soportes sólidos insolubles en la síntesis orgánica. Actualmente sus líneas de trabajo se centran en la síntesis de butenolidas y en la aplicación de la fase sólida a reacciones fotoquímicas. Ha sido vicerrector de investigación de la UAB.

SIDNEY PESTKA ("Purificación y fabricación de interferones humanos") dirige el laboratorio de genética molecular del Instituto Roche de Biología Molecular. De origen polaco, se formó en los Estados Unidos. Comenzó sus estudios en la Universidad de Princeton y se doctoró en medicina por la de Pennsylvania. Médico residente en las secciones de pediatría y en la de medicina general del hospital municipal de Baltimore, ingresó luego en el Instituto Nacional de la Salud. Trabajó en el departamento de cardiología y oncología de dicha institución hasta que entró, en 1969, en la Roche. Pestka es profesor adjunto de patología de la Facultad de Medicina de la Universidad de Cambridge. A él se debe gran parte del trabajo básico que resultaría en la purificación de los interferones.

E. N. PARKER ("Campos magnéticos en el cosmos") es catedrático de física, astronomía y astrofísica de la Universidad de Chicago. Inició su carrera en la Universidad del estado de Michigan, concluyéndola con el doctorado en física en el Instituto de Tecnología de California, en 1951. Los cuatro años siguientes los pasó en la Universidad de Utah, primero en el departamento de matemática y en el de física luego. Se trasladó a Chicago en 1955. En su opinión, "las observaciones astrofísicas crean múltiples rompecabezas de física clásica, con pistas para descubrir los efectos físicos, hasta ahora desconocidos, que los producen. Algunos de nosotros, de talante escéptico, vivimos desenmascarando explicaciones hueras [de tales fenómenos] y tratando de buscar alguna más sólida. Y ello nos lleva al descubrimiento de muchos efectos pequeños, ignorados hasta ahora, que se van acumulando a lo largo de los años y logran ordenar algunos de los rompecabezas principales".

ROY S. LEWIS y EDWARD ANDERS ("Materia interestelar en meteoritos") forman parte del claustro docente de la Universidad de Chicago, unidos por un interés común en los aspectos químicos de la astronomía. Lewis tiene plaza contratada de investigador. Se licenció en Berkeley, en 1967, trasladándose a Chicago cinco años más tarde, poco antes de que la Universidad de California en Berkeley le concediera el doctorado en ciencias atmosféricas y espaciales. "Me siento atraído, cuenta, por la historia primitiva del sistema solar y, de un modo general, por la historia de la nucleosíntesis en el universo. He contribuido en ese campo de la ciencia con la medición de la composición isotópica y la abundancia de los gases nobles aprisionados en los meteoritos primitivos." Anders es titular de la cátedra Horace B. Horton de química. Está en Chicago desde 1955. Letón de nacimiento, cursó su primer ciclo universitario en Munich, de 1946 a 1949. Se doctoró en radioquímica, en 1954, por la Universidad de Columbia.

GLENN D. PRESTWICH ("Las defensas químicas de los termes") es profesor asociado de química de la Universidad estatal de Nueva York en Stony Brook. Se licenció por el Instituto de Tecnología de California en 1970 y, en 1974, se doctoró en química por la Universidad de Stanford. Antes de incorporarse a Stony Brook trabajó durante tres años como investigador en la Universidad de Cornell y en el Centro Internacional de Fisiología y Ecología de los Insectos de Kenia. Ha recibido una beca de la Fundación Alfred P. Sloan, para el bienio 1981-83, por su labor en la química de sustancias sintetizadas por insectos.

DOUGLAS A. BLAIR y ROBERT A. POLLAK ("Decisiones racionales colectivas") son economistas, cuyos campos de interés respectivo se intersecan en el tema de su artículo. Blair es profesor asociado de economía de la Universidad de Rutgers. Tras licenciarse por el Swarthmore College (1970) se doctoró en economía por la Universidad de Yale, en 1976. Ese mismo año ingresó en el claustro docente de la Universidad de Pennsylvania, donde permaneció hasta 1981, cuando pasó a Rutgers. Pollak es titular de la cátedra Charles y William Day, de economía y ciencias sociales, de Pennsylvania. Asistió al Amherst College, por el que se licenció en 1960. Se recibió de doctor en economía por el Instituto de Tecnología de Massachusetts, en 1964, año en que ingresó en el claustro de Pennsylvania, donde ha permanecido desde entonces.

PETTER AUNE, RONALD L. SACK y ARNE SELBERG ("Iglesias noruegas de pies derechos") son ingenieros expertos en estructuras antiguas de madera. Aune es catedrático de estructuras lígneas en el Instituto Noruego de Tecnología (NTH) de Trondheim. Se formó en ingeniería civil en el NTH. Sack es profesor de ingeniería civil en la Universidad de Idaho. Cursó estudios en la Universidad de Minnesota, por la que se doctoró en 1964. Comenzó a trabajar en iglesias de pies derechos en 1976 y 1977, coincidiendo con un año sabático que, becado por el Real Consejo Noruego para las Ciencias y la Investigación Industrial, pasó en el NTH. Hasta su jubilación, Selberg dirigió la sección de estructuras de acero del NTH. Sus trabajos se han centrado en puentes colgantes.

CHARLES BIGELOW y DONALD DAY ("Tipografía digital") comparten una común afición por la forma de las letras y los ordenadores. Charles Bigelow es profesor ayudante de ordenadores y de creación artística por ordenador en la Universidad de Stanford. En 1967 se licenció por el Reed College. En 1982 recibió una beca de la Fundación MacArthur por sus contribuciones al estudio de la tipografía. Con su colega Kris Holmes, diseña tipos de letras para sistemas de ordenador. Donald Day es profesor del departamento de vídeo del Colegio de Artes y Oficios de California. Como Bigelow, estudió en el Reed College. Desde 1979 colabora con un grupo dirigido por Daniel Sandin, de la Universidad de Illinois, que trabaja en el diseño de procesadores de imágenes digitales.

Hace...

José M.ª López Piñero

... cien años

Falleció en Madrid Francisco Méndez Alvaro. Había nacido en 1806 en la localidad abulense de Pajares y estudió cirugía en el Colegio de San Carlos de Madrid hasta 1828, fecha en la que esta institución pasó a convertirse en Colegio de Medicina y Cirugía. La reforma perjudicó a Méndez Alvaro, que tuvo que licenciarse primero en cirugía y luego en medicina, recibiendo, por último, el título de doctor en medicina y cirugía (1834). Ejerció a continuación durante tres años como médico rural y también como cirujano militar. En el ejército fue ayudante del higienista Mateo Seoane, que acababa de regresar de su exilio en Londres y cuyas ideas influyeron decisivamente en su posterior orientación.

En marzo de 1837 se trasladó a Madrid y abandonó la práctica clínica, tanto hospitalaria como privada. Se dedicó al periodismo, dirigiendo o colaborando en diferentes diarios y revistas de información general, y participó en la vida política desde el partido moderado. Al mismo tiempo, desarrolló una intensa actividad como higienista y, sobre todo, como promotor y redactor de publicaciones médicas.

Méndez Alvaro llegó a ocupar puestos de gran influencia en las instituciones sanitarias oficiales. Recordaremos únicamente que fue el principal responsable de la reforma hospitalaria madrileña de 1845, siendo presidente de la Junta Municipal de Beneficencia, y que perteneció al Consejo de Sanidad del Reino a partir de su fundación en 1847. A través de Seoane, asimiló la concepción liberal de la higiene pública en su vertiente británica, pero la entendió desde una mentalidad política estrictamente conservadora. Por ello, la evolución de su actitud ante los nuevos problemas que planteaba el proletariado industrial contrasta con la apertura que siempre mantuvo Pedro Felipe Monlau, el otro gran discípulo de Seoane. Dicha evolución se refleja de modo muy expresivo en dos trabajos suyos: *Consideraciones sobre la higiene pública y mejoras que reclama en España la higiene municipal* (1853) y *Estudio higiénico-social de la habitación del pobre* (1874). El contenido del primero puede dividirse en tres partes: la primera da un excelente informe del movimiento de higiene pública en la Europa contemporánea, principalmente en Alemania, Francia, Inglaterra y Bélgica; la segunda analiza los factores del atraso higiénico español, denunciando, entre otras deficiencias, el estado lamentable de las viviendas proletarias, lo frecuente de la adulteración de los alimentos y la carencia de estudios estadísticos; la tercera parte propone las mejoras que reclama en España la higiene municipal para superar tal atraso. Méndez Alvaro demuestra conocer con rigor las corrientes europeas relacionadas con la salud pública. Muy significativo es, por ejemplo, su interés por la obra de Chadwick, así como la importancia que concede a las estadísticas demostrativas del descenso de vida media en las ciudades industriales inglesas como punto de partida de todo un movimiento. Méndez Alvaro concede mucho peso en la aparición de la gran mortalidad de las clases pobres en los grandes centros urbanos a las malas condiciones de sus viviendas. Denuncia, por ello, su estado: "Todo el mundo conoce la precipitación con que las casas se edifican, sobre todo en Madrid; nadie ignora que se emplean malos materiales para invertir el menor capital posible y que, ávidos los propietarios de un interés crecido, aprovechan el terreno construyendo habitaciones para muchas familias donde apenas hay espacio para una sola. Este mal, que cada día va en aumento, reclama muy pronto y radical remedio, porque no ha de posponerse la salud pública al interés individual, consintiendo que sigan construyéndose patios pequeños y oscuros, especie de tubos por donde no penetran ni el aire ni la luz, y que se multiplique el número de los cuartos haciendo habitaciones mezquinas, lóbregas, insuficientes para las pobres familias que las habitan... Nada hay, en efecto, tan sombrío, tan triste e insalubre como esos reducidos albergues de la miseria, esas habitaciones inmundas donde el pobre vive, crece, se deteriora y muere casi sin aire para respirar, sin

Año I. 1.° de Enero de 1854. Número 1.°

En Madrid 12 rs. el trimestre.
Redaccion, pretil de los Consejos, número 3.
En provincias 15 rs. el trimestre.
En casa de los comisionados ó mediante libranzas.

EL SIGLO MÉDICO.

Ventajas para los suscritores.
Pueden tomar las obras publicadas en la *Biblioteca de medicina* y *Museo científico*, con la rebaja de un 10 por 100 de sus precios.

(BOLETIN DE MEDICINA Y GACETA MÉDICA.)

PERIÓDICO OFICIAL DE LA REAL ACADEMIA DE MEDICINA DE MADRID Y DE LA SOCIEDAD DE SOCORROS MÚTUOS.

RESUMEN.

ESCRITOS ORIGINALES: FILOSOFIA MEDICA: Rápida reseña de su estado actual.—MEDICINA ADMINISTRATIVA: ¿Qué reforma exige la ordenanza de 1804?—BOLETIN CLINICO: Hospitales generales de Madrid.—ASUNTOS PROFESIONALES: Títulos falsos ó falseados.—PRENSA MEDICA. Anatomía y fisiologia: Medicina.—PARTE OFICIAL: Disposiciones del Gobierno.—SANIDAD MILITAR: Reales órdenes.—SOCIEDAD MEDICA GENERAL DE SOCORROS MUTUOS.—VARIEDADES: Premios á la limpieza.—Enfermedad sospechosa en Galicia.—Oposiciones á baños.—GACETA DE EPIDEMIAS.—CRONICA.—VACANTES.

un esqueleto el sexo, la edad y otras muchas circunstancias del sugeto á quien perteneció. Asi es que la medicina ha participado siempre del espíritu filosófico de las diversas épocas, siguiendo paso á paso los de los varios sistemas dominantes. Sin embargo, aunque esto en general sea una verdad indudable, necesita algunas aclaraciones.

No ha de creerse que la medicina ha seguido siempre servilmente las indicaciones de la filosofia. Aunque sujeta al movimiento general, ha gimetafísica y las altas especulaciones de la filosofia un escalon por debajo de la enagenacion mental.

Y tienen motivo, mas no razon, para pensar asi. Tienen motivo, porque la filosofia por desgracia ha caido harto á menudo en las aberraciones mas estrañas; porque es facil abusar de ella volviéndola en contra hasta del sentido comun; porque sus pretensiones han sido tal vez demasiado ambiciosas, y porque en la mayor par-

1. Cabecera del primer número de **El Siglo Médico**, *revista fundada por Méndez Alvaro, que continuaría publicándose hasta 1947. Se la considera generalmente como la más importante revista médica general de la historia del periodismo médico español*

luz, helado de frío unas veces, abrasado de calor otras, confundido con los animales domésticos, aspirando los gases de las letrinas o de las inmundicias, mezclados los sexos y las edades, sobre un suelo tal vez húmedo, con las paredes ennegrecidas y sucias, sin otros muebles que algún trasto viejo e inútil, sin más abrigo que unos cuantos andrajos". Para solucionar tan lamentable situación propone la construcción de casas y barrios para obreros como los de los demás países, informando de la real orden de aquel mismo año, a la que critica de "vaga y confusa".

El punto de partida del trabajo de Méndez Alvaro de 1874 es muy semejante al anterior: insiste en la mayor mortalidad existente en las grandes ciudades en relación con las pequeñas y con el campo, utilizando las estadísticas de Farr, Stark, Tourtelle y otros autores. Sin embargo explica tal fenómeno de modo más complejo. La aglomeración, la existencia de sustancia orgánica en estado de putrefacción, la insalubridad de las habitaciones proletarias, la dañosa influencia de las industrias y la desmoralización se consideran como causas principales y, por encima de todo, la miseria, a la que consagra un fino análisis. Vuelve a utilizar las estadísticas ya clásicas sobre la desigualdad de la vida media entre ricos y pobres, y añade otras recientes; cita a Malthus, que "aunque no en todo puedan admitirse sus doctrinas... y menos algunas de las consecuencias que de ellas suelen deducirse, al menos es indisputable (su) ecuación entre la cantidad de la vida y los alimentos que la sostienen"; llega incluso a incluir buena información acerca del aumento de la mortalidad infantil en las poblaciones fabriles. Sigue pensando en la conveniencia de construir casas para obreros, pero no con el entusiasmo de 1853, sino "dentro de determinados y prudentes límites, con grande parsimonia y extremada cautela, favoreciéndola indirectamente por medio de leyes y disposiciones municipales bien entendidas; de ninguna manera dándola el carácter socialista que se la ha dado en Francia durante el Imperio de Napoleón III", régimen que critica ferozmente. Las limitaciones de su ideología liberal se ponen bien de manifiesto cuando afirma que el estado "no puede meterse a constructor, ni aun a subvencionar la construcción de obras semejantes, destinando al efecto cantidades más o menos crecidas del Tesoro público, si es que voluntariamente no se arroja en brazos del socialismo, que cada día se haría más exigente e imperioso". Méndez Alvaro, impresionado por "las perniciosas doc-

CONSIDERACIONES

SOBRE LA

HIGIENE PÚBLICA,

Y MEJORAS QUE RECLAMA EN ESPAÑA LA

HIGIENE MUNICIPAL.

MEMORIA PRESENTADA

A LA REAL ACADEMIA DE MEDICINA DE MADRID,

PARA SU ADMISION COMO SOCIO DE NÚMERO,

POR DON FRANCISCO MENDEZ ÁLVARO,

DOCTOR EN MEDICINA Y CIRUGIA DE LA FACULTAD DE ESTA CÓRTE, SECRETARIO HONORARIO DE S. M., COMENDADOR DE LA REAL Y DISTINGUIDA ÓRDEN ESPAÑOLA DE CARLOS III, CONDECORADO CON LA CRUZ DE EPIDEMIAS, VOCAL QUE HA SIDO DE LA JUNTA PROVINCIAL DE SANIDAD DE MADRID Y DE LA MUNICIPAL DE BENEFICENCIA, SOCIO CORRESPONSAL DE LAS REALES ACADEMIAS DE MEDICINA DE SEVILLA Y GRANADA, DE LA SOCIEDAD MÉDICA DE LONDRES, DE LA DE MEDICINA PRÁCTICA DE PARIS, DE LA FÍSICO-MÉDICA DE ERLANGEN, DE LA DE CIENCIAS MÉDICO-NATURALES DE DRESDE Y DE VARIAS OTRAS NACIONALES Y ESTRANJERAS, ANTIGUO AYUDANTE DEL CUERPO DE SANIDAD MILITAR, SECRETARIO DEL CONSEJO DE SANIDAD DEL REINO, ETC., ETC.

«Cuando los esfuerzos individuales no alcanzan á conjurar el mal, debe intervenir el Estado de una manera activa.»
(P. FRANCK.)



MADRID.

IMPRENTA Á CARGO DE JOSÉ RODRIGUEZ, CALLE DE S. VICENTE BAJA, NÚM. 74.

1853.

2. En sus **Consideraciones sobre la higiene pública** *(1853), Méndez Alvaro se enfrentó con los problemas sanitarios del proletariado urbano desde los supuestos de la higiene liberal, sin los elementos reaccionarios que su ideología política moderada introdujo en su obra médicosocial tardía*

trinas socialistas y demagógicas que tanto cuerpo han tomado en nuestros días", no duda en criticar una tarea al parecer tan indiscutible como la construcción de viviendas modestas, a la que llega a calificar de propias del "pensamiento socialista y un paso muy avanzado hacia el comunismo".

Entre los demás estudios de Méndez Alvaro sobre temas médicosociales destacan los titulados *La lepra en la España a mediados del siglo XIX* (1860) –en el que analiza los resultados de la estadística oficial de 1851-52– y *Discurso acerca de la preservación de las viruelas* (1871).

Más importante todavía fue la contribución de Méndez Alvaro al desarrollo en España de las editoras y el periodismo médicos modernos. De sus numerosas iniciativas en este terreno, recordaremos, en primer término, la fundación, en 1854, de la revista *El Siglo Médico*, por fusión del *Boletín de Medicina y Cirugía y Farmacia*, que Méndez Alvaro dirigía desde 1847, con la *Gaceta Médica* de Matías Nieto Serrano. *El Siglo Médico*, que se publicó hasta 1947, es generalmente considerado como la revista médica general de mayor relieve de la historia del periodismo médico español. Fue también uno de los directores de la *Biblioteca Escogida de Medicina y Cirugía* (1841-51), colección que publicó en poco más de una década más de treinta obras, casi todas ellas traducciones. Cultivó asimismo la bibliografía médica.

Química en fase sólida

La mayoría de los procesos de laboratorio se basan en colisiones al azar de moléculas que se mueven libremente. Se introduce aquí un nuevo método, en el que las moléculas se fijan a un soporte insoluble

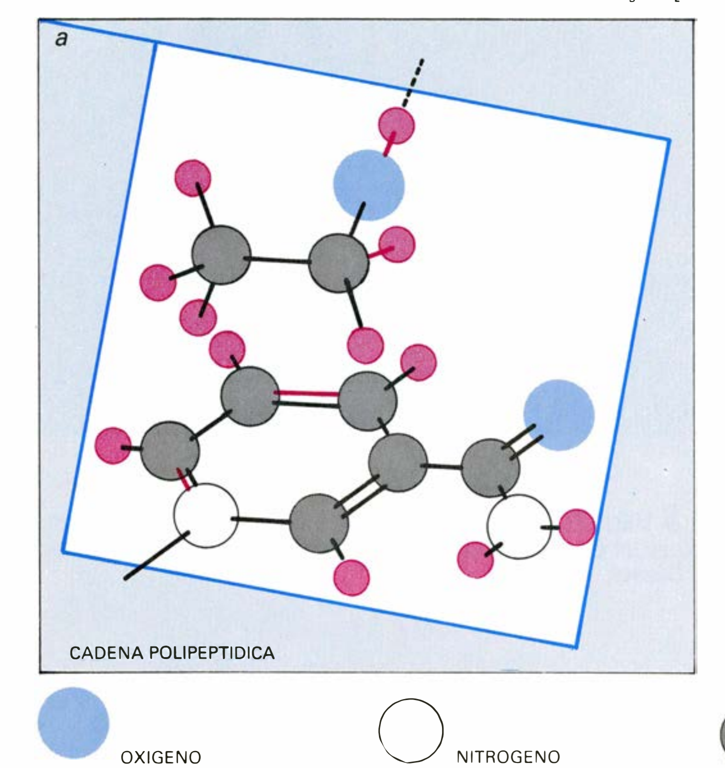
Josep Font

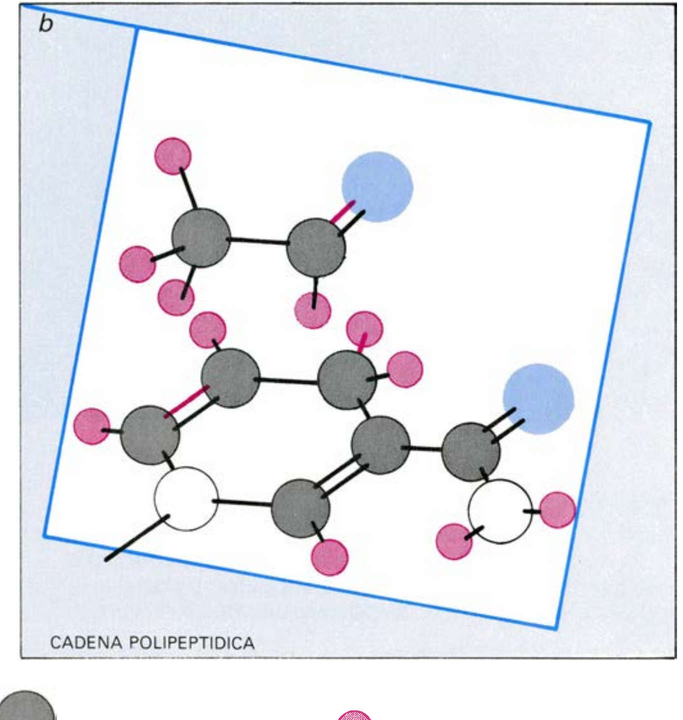
La teoría atómica de la materia en general y la teoría cinética de los fluidos en particular pertenecen ya al lenguaje ordinario. En ese mismo tenor, la expresión reacción química evoca inmediatamente el choque entre dos especies químicas (A y B) que, a través de un reordenamiento de sus átomos o simplemente por intercambio de electrones, se convierten en nuevas especies ($C + D$). Reacción que se asocia, al propio tiempo, al choque entre especies que se mueven libremente bien en fase gaseosa, bien en fase líquida donde las moléculas reaccionantes se hallan disueltas. Si esta disolución es diluida, su comportamiento puede explicarse cual si se tratara de las moléculas de un gas. En todo caso, estamos imaginando reacciones químicas en fase homogénea, es decir, en situaciones en las que cualquier elemento o parte de volumen de la fase gaseosa o de la fase líquida representa exactamente la composición y las propiedades del conjunto del medio de reacción.

El tipo de reacción al que nos venimos refiriendo constituye una situación límite explotada por los químicos en sus laboratorios para explicar, en el microuniverso del matraz, parte de las reacciones químicas que tienen lugar en el universo. Situación que ha sido espectacularmente aprovechada por la industria para sintetizar —generalmente a través de reacciones en fase homogénea— miles y miles de nuevos materiales.

En la naturaleza, donde también se dan reacciones en fase homogénea (en la atmósfera, en el mar y en los ríos), predominan las reacciones en fase heterogénea: las que se producen entre dos fases no miscibles (por ejemplo sólido-líquido o líquido-líquido). Los choques intermoleculares se producen aquí sólo en la interfase. (El elemento de volumen del conjunto reaccionante no representa ya exactamente la composición y las propiedades de todo el medio reaccionante.) Para entender la evolución de la reacción química en fase heterogénea debe conocerse con detalle la composición y la estructura tridimen-

$NADP^+ + CH_3{-}CH_2OH \rightarrow NADPH + CH_3{-}CHO + H^+$





OXIGENO NITROGENO CARBONO HIDROGENO

1. ACCION ENZIMATICA, en este caso oxidación del etanol a acetaldehído en presencia del enzima deshidrogenasa (*fondo azul*) que posee el cofactor $NADP^+$. Uno, y sólo uno, de los hidrógenos del etanol, el que posee la unión simbolizada en rojo en forma de ion hidruro, ataca al catión piridino del cofactor, convirtiendo el $NADP^+$ en NADPH. La parte proteica del enzima acepta el hidrógeno hidroxílico en forma de protón: como consecuencia, el etanol se oxida a acetaldehído. La oxidación tiene lugar porque el enzima acerca de modo conveniente la molécula de etanol hasta el cofactor. (En *a*, la molécula superior corresponde al etanol; la inferior, al cofactor; en *b*, la molécula superior corresponde al acetaldehído; la molécula inferior, al cofactor.)

sional de la interfase y del lugar concreto en donde la reacción tiene lugar.

Ejemplos de reacciones químicas en fase heterogénea son: (*a*) las que tienen lugar en los suelos (arcillas), donde los silicatos actúan de plantilla o molde para determinados intercambios de cationes, o bien de catalizadores para determinadas transformaciones químicas de sustancias orgánicas; y (*b*) las que tienen lugar en las células vivas, en donde el protoplasma, con sus orgánulos, no constituye una solución homogénea de sustancias, sino que forma toda una sucesión de heterogeneidades en donde los enzimas —proteínas de peso molecular generalmente elevado— inducen reacciones químicas específicas, en puntos determinados del sistema celular.

Recientes teorías sobre el origen de los seres vivos en la Tierra confieren un papel importante a la acción catalizadora heterogénea de las arcillas u otros tipos de suelo. Por otra parte, el hombre no ha hecho más que imitar, y de modo muy imperfecto, las posibilidades sintéticas de la naturaleza. Mediante sus sistemas enzimáticos, ésta es capaz de aprovechar la energía luminosa, transformando el dióxido de carbono en carbohidratos (fotosíntesis), y efectuar transformaciones químicas que conducen a moléculas tan complejas como la estricnina o la hemoglobina.

Del estudio de las reacciones químicas en heterofase, como las citadas, se ocupan los químico-físicos dedicados a fenómenos de superficie y los bioquímicos enzimólogos. Siguiendo los pasos de estos últimos, los químicos sintéticos utilizan los modos y métodos de la heterofase para intentar aprovecharse de algunas de sus ventajas y así imitar mejor a la naturaleza. Aquí nos vamos a centrar en un aspecto de la síntesis orgánica en fase heterogénea que emplea polímeros como soporte de reactivos. Ello nos permitirá establecer una correlación entre estos métodos artificiales y el que usan los enzimas. (Téngase presente que los enzimas son polímeros proteicos que soportan centros activos donde tienen lugar la reacción de transformación química del substrato.)

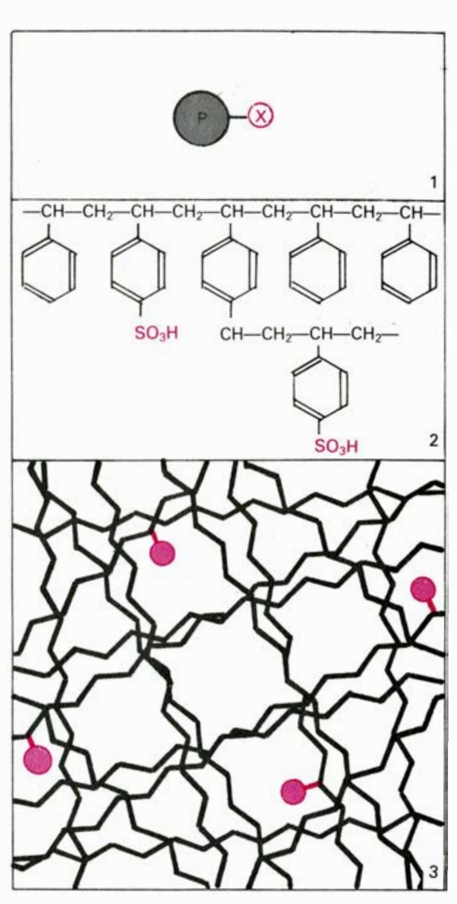
¿Y cuál es el modo de acción de los enzimas? Un enzima es una proteína catalítica, y como toda proteína, formada por aminoácidos. Suele poseer una naturaleza globular: la cadena polipeptídica se enrolla sobre sí misma, confiriendo a la molécula una forma redondeada. Esta estructura tridimensional y la presencia de un grupo prostético son los causantes de la acción catalítica de la proteína. El grupo prostético o cofactor es una molécula no proteínica unida covalentemente a la cadena polipeptídica (aunque también puede hallarse unida mediante otras fuerzas), que efectúa la transformación química del sustrato, en condiciones mucho más favorables de reacción que si ésta tuviera lugar en fase homogénea: la energía de activación de la reacción disminuye y de ahí que se diga que la reacción se halla catalizada.

Los enzimas catalizan los mismos tipos de reacciones que solemos encontrar en cualquier laboratorio de química orgánica: oxidaciones, reducciones, hidrólisis, hidroxilaciones, eliminaciones, alquilaciones, etcétera.

Relacionadas con su modo de acción, los enzimas gozan de ciertas propiedades características, cuyas dos principales son: aumentar la velocidad de las reacciones catalizadas en unas 10^9 - 10^{12} veces con respecto a las mismas reacciones efectuadas con métodos convencionales, y, en segundo lugar, actuar de modo muy específico. En efecto, cada enzima suele catalizar un tipo particular de reacción y acepta un determinado substrato. Esta especificidad llega a extremos como el de discriminar un punto especial de la molécula del sustrato en donde se produce la transformación (regioselectividad) o el de aceptar sólo uno de los enantiómeros de una mezcla racémica (la diferencia entre dos moléculas enantioméricas es la misma que existe entre la mano derecha y la izquierda).

El efecto cinético (rebajar la energía de activación) del enzima y su especificidad se deben, probablemente, a la disposición tridimensional del sustrato y a los cofactores en el centro o punto activo del enzima. El sustrato sólo puede "introducirse" en este centro activo en una posición determinada, adquiriendo, junto al cofactor, una conformación muy parecida a la del estado de transición de la reacción química transformadora.

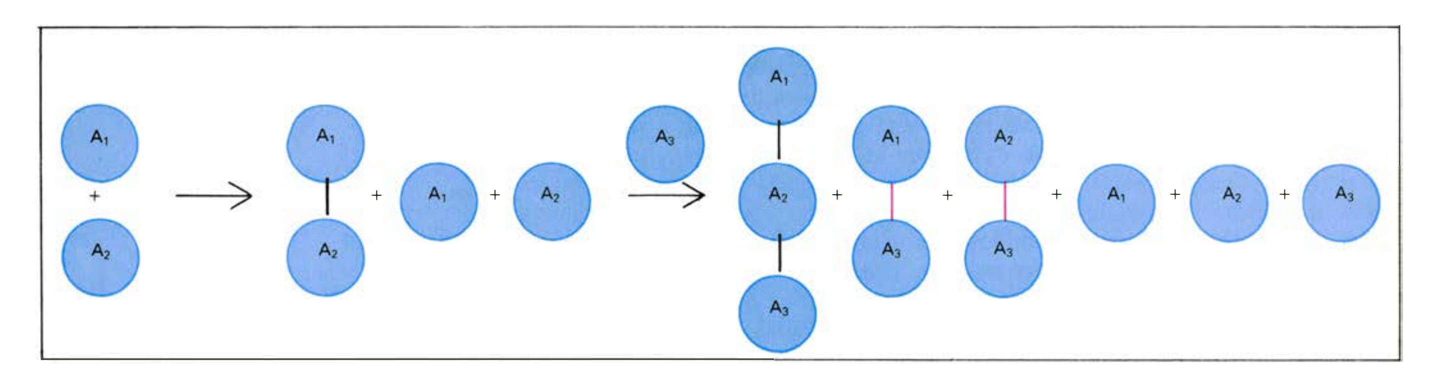
Un ejemplo aclarará y fijará conceptos. Las reductasas o deshidrogenasas son enzimas que catalizan reacciones de oxidación-reducción, que transcurren con transferencia de un ion hidruro. Estos enzimas poseen un grupo prostético o cofactor, el nicotinamida adenina dinucleótido (NADPH). El enzima, plegado, deja expuestas dos zonas muy distintas: en una está fijado el cofactor y por la otra se "une" el substrato, etanol. En el proceso de oxidación con $NADP^+$, de etanol a acetaldehído, la parte activa del cofactor, la molécula de nicotinamida convenientemente cuaternizada (en forma por tanto catiónica) acepta un ion hidruro de la molécula de etanol, y precisamente de uno sólo de los hidrógenos unidos al átomo de carbono oxidado.



2. LAS RESINAS de intercambio iónico constituyen un ejemplo del uso de polímeros con grupos funcionales. En *1*, se simboliza la resina, siendo *X* el grupo funcional. En *2* se ilustra la estructura química de una resina sulfónica: poliestireno, entrecruzado o copolimerizado con divinilbenceno, y funcionalizado con el grupo sulfónico ($-SO_3H$). En este caso, la resina intercambia cationes. En *3* se representa la estructura tridimensional de un copolímero, en la que se indican los puntos donde está anclado un determinado reactivo (*círculos rojos*). Si el copolímero es de estireno y divinilbenceno, las cadenas significan el esqueleto hidrocarbonado del que penden los anillos bencénicos (no dibujados). Para resinas de intercambio iónico, el reactivo sería $-SO_3H$, $-COOH$, $N^+(CH_3)_3$, $-CH_2 - N^+(CH_3)_3$, etcétera. En el polímero utilizado por Merrifield para la síntesis secuenciada de péptidos en fase sólida, el reactivo era $-CH_2Cl$.

Así actúan los enzimas en la naturaleza. Durante los diez o quince últimos años la idea de utilizar polímeros en los que se uniera o anclara un reactivo químico, biomimetizando la acción de los enzimas, ha prendido con fuerza entre los químicos sintéticos. Pero el pionero en el aprovechamiento de esa técnica, R. B. Merrifield, andaba buscando otra ventaja: la insolubilidad del soporte polimérico, con lo que se podía lavar fácilmente y, por tanto liberar, el medio de reacción de reactivos que pudieran molestar posteriormente.

La verdad es que no era desconocido el uso de polímeros con grupos funcionales reactivos. Piénsese en las resinas



3. SINTESIS SECUENCIADA DE UN TRIPEPTIDO. La primera reacción forma el dipéptido A_1-A_2, pero en la metodología clásica resultaba muy difícil separar dicho dipéptido de las impurezas de los aminoácidos iniciales, A_1 y A_2. En virtud de ello, la formación del tripéptido A_1-A_2-A_3, que evidentemente constituye el producto principal de la síntesis, se ve alterado con la presencia de dipéptidos de secuencia errónea, sobre todo el formado por A_1-A_3.

de intercambio iónico. Una resina de intercambio iónico no es más que un polímero insoluble conseguido por copolimerización de estireno con una pequeña cantidad de divinilbenceno. Resulta así un copolímero entrecruzado o tridimensional bastante inerte desde el punto de vista químico, pues está constituido por una cadena hidrocarbonada saturada de la que penden (en átomos de carbono alternados) anillos bencénicos. Estos anillos bencénicos pueden ser funcionalizados, o lo que es lo mismo, pueden sulfonarse discrecionalmente, convirtiendo el polímero en un ácido bencenosulfónico "polimérico". [Con ello pretendemos decir que algunos (sea uno de cada 20) de los anillos bencénicos del poliestireno, entrecruzados con divinilbenceno, contienen el grupo sulfónico ($-SO_3H$).] Las propiedades de este ácido bencenosulfónico "polimérico" son algo distintas del ácido bencenosulfónico monómero. Este es relativamente soluble en agua, donde actúa como ácido fuerte; aquél es insoluble en agua, aunque absorbe en su malla polimérica cierta cantidad de agua produciéndose un hinchamiento y actúa como ácido liberando protones o intercambiándolos por cationes, como el sodio. En forma de sal sódica intercambia lentamente este catión fijado sobre la malla polielectrolítica polimérica por otros cationes que se introduzcan en el medio. En este principio se fundan las resinas intercambiadoras de cationes, cuyo grupo reactivo anclado sobre soporte sólido polimérico es el grupo sulfónico. Cuando el grupo anclado sobre el mismo soporte es el grupo carboxilo (COOH, un grupo ácido débil) o cuando el grupo es el trimetilamonio ($N(CH_3)_3$), las resinas correspondientes realizan entonces funciones de intercambiadoras de aniones.

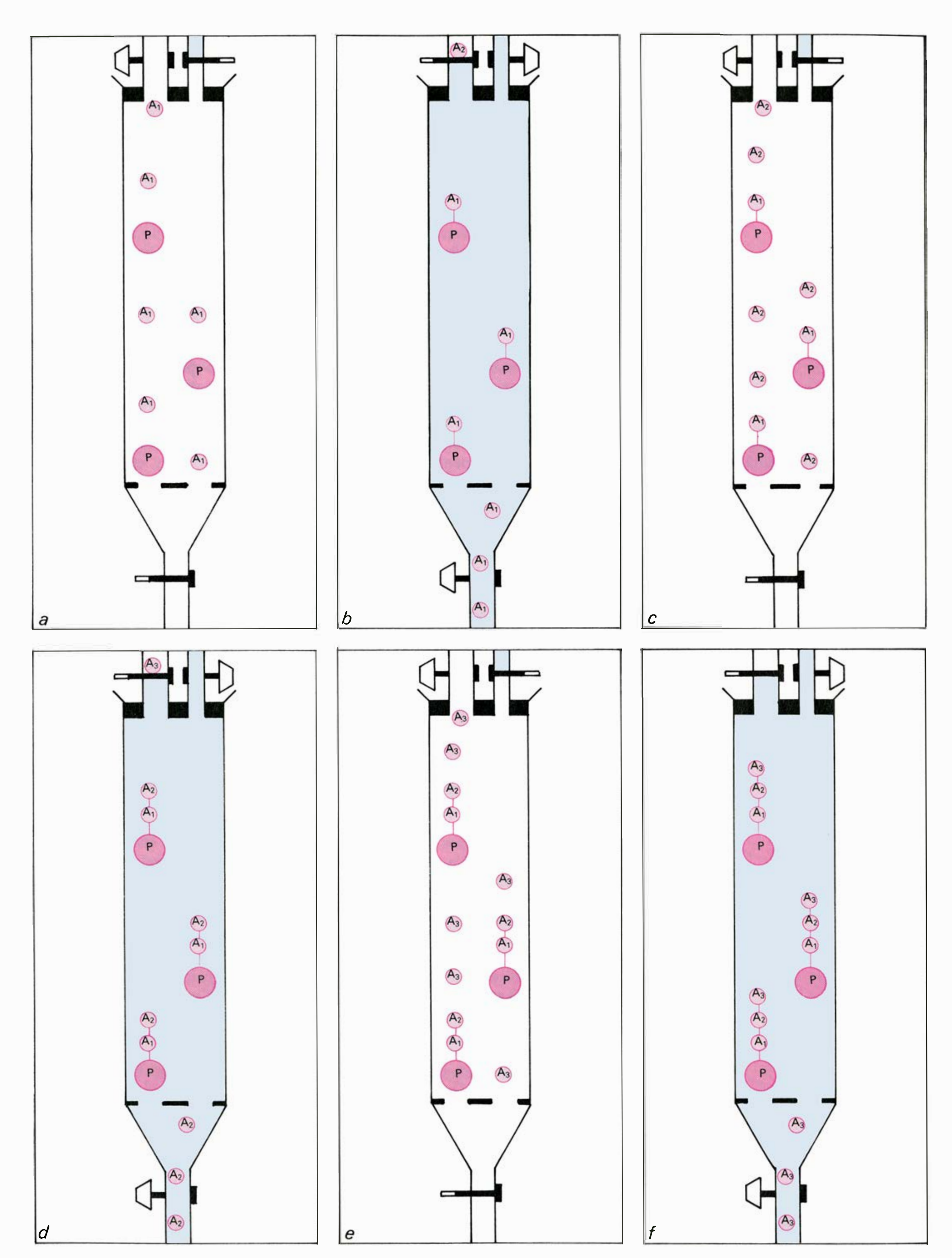
Merrifield publicó en 1963 su primer trabajo en el que describía el uso de un polímero sólido insoluble —también poliestireno entrecruzado con un 2 por ciento de divinilbenceno— en la síntesis secuenciada de un polipéptido. Su idea fue realmente brillante. En la síntesis de laboratorio de un polipéptido es preciso concatenar secuenciadamente los distintos aminoácidos, llamémosles A_1, A_2 y A_3; la primera reacción de formación de enlace peptídico no ofrece especiales problemas y se logra obtener el dipéptido A_1-A_2, aunque impurificado por restos aminoácidos de A_1 y A_2. Cuando el aminoácido A_3 se hace reaccionar con el dipéptido A_1-A_2 se obtiene el tripéptido A_1-A_2-A_3 y, también, pequeñas cantidades reducidas de A_1-A_3 y A_2-A_3. Es decir, que el tripéptido objetivo A_1-A_2-A_3 viene impurificado con A_1-A_3 y A_2-A_3 (que no poseen la secuencia adecuada) junto con los aminoácidos de partida A_1 y A_2. Es fácil deducir que en la síntesis de un polipéptido tan pequeño como la oxitocina (con nueve aminoácidos) las posibilidades de secuencias erróneas son enormes y, por tanto, el péptido final aparecerá muy impurificado. Aunque existen métodos y técnicas de separación de los péptidos, los procedimientos son tediosos, laboriosos y largos.

Merrifield decidió anclar —léase unir covalentemente— el primer aminoácido a un polímero funcionalizado o activo: el poli (4-clorometilestireno). Este polímero no es más que el poliestireno entrecruzado, que llamaremos P, al que, mediante una reacción de clorometilación, se agrega el grupo clorometilo ($-CH_2Cl$), al azar y estadísticamente, en uno de cada quince anillos bencénicos. El grupo clorometilo representa la funcionalización o centro activo del polímero, ya que el esqueleto feniletílico es relativamente inerte, al menos frente a los reactivos con que será tratado el polímero. En este centro activo se ancla el primer aminoácido a través de un grupo amino (H_2N-) del aminoácido mediante una reacción de sustitución: se logra así un polímero, P $-CH_2$-A_1, que contiene este aminoácido fijado sobre el polímero, libre de interacciones intermoleculares consigo mismo (puesto que sus moléculas se hallan muy alejadas unas de otras); viene a ser como si las moléculas del aminoácido A_1 estuvieran congeladas sobre la fase sólida polimérica, cual cerdas de un cepillo.

El aminoácido A_2 puede entonces reaccionar con P $-CH_2A_1$ pero con la sustancial ventaja de que este polímero —recuérdese, insoluble— puede ser lavado exhaustivamente para liberarlo de la presencia de cualquier molécula de A_1 o A_2 no ligados al polímero. Se forma así P $-CH_2$-A_1-A_2. La iteración de las operaciones de formación de enlace peptídico con los aminoácidos previstos en la secuencia y de lavado después de cada reacción de unión de aminoácido produce el crecimiento del polipéptido anclado sobre el polímero, es decir, a P $-CH_2$-A_1-A_2-A_3... A_n. El polipéptido se libera de la unión covalente con el grupo CH_2 mediante una reacción de escisión con ácido fluorhídrico.

Con esta técnica, Merrifield conseguía péptidos muy puros. Más aún, concentraba la formación de péptidos sobre una superficie sólida sin menoscabo del rendimiento sintético con purificaciones intermedias. El resultado fue extraordinariamente esperanzador y su técnica ha significado una ruptura o avance decisivo en la síntesis de polipéptidos con respecto a los métodos seguidos hasta entonces. El propio Merrifield, junto con Gutte, anunció en 1969 la síntesis total del enzima ribonucleasa *A* por este método. Dicha proteína posee una secuencia de 124 aminoácidos.

El método de la fase sólida se ha extendido ya a la síntesis de otros biopolímeros: polisacáridos y polinucleótidos. Demos un paso adelante para ocuparnos de una nueva dimensión de los polímeros funcionalizados: su utilización



4. METODO DE SINTESIS SECUENCIADA de un tripéptido en fase sólida. En la ilustración, *P* representa un copolímero de estireno y divinilbenceno. Las etapas son las siguientes: *a*, introducción del aminoácido A_1 en la columna; *b*, fijación de A_1 al copolímero y eliminación mediante lavado de los restos de A_1; *c*, introducción del aminoácido A_2 en la columna; *d*, formación del dipéptido A_1-A_2 y eliminación, mediante lavado, de los restos de A_2; *e*, introdución del aminoácido A_3 en la columna; *f*, formación del tripéptido deseado A_1-A_2-A_3 y eliminación mediante lavado de los restos del aminoácido A_3.

5. OXIDACION DE UNA OLEFINA, en este caso el tetrametiletileno, con un perácido polimérico.

en la síntesis de moléculas orgánicas, en donde pueden ayudar a resolver algunos problemas sintéticos.

Las líneas fundamentales de esta técnica recuerdan la síntesis de tipo secuencial, repetitiva, de polipéptidos: un soporte inerte donde anclar moléculas que, en el momento de la reacción química, se hallan "congeladas" y relativamente diluidas en la trama polimérica, en una situación que puede denominarse hiperentrópica (de desorden máximo). Al soporte sólido se une un catalizador, un reactivo o un substrato; la síntesis de la molécula-objetivo puede realizarse permanentemente sobre el soporte o bien utilizarse éste sólo en algún paso de la síntesis

Por anclado o ligado designamos la situación en que moléculas orgánicas de estructura bien definida se hallan unidas covalentemente al soporte polimérico. El término resina remite, por lo común, a un poliestireno entrecruzado o copolimerizado con divinilbenceno (1 o 2 por ciento) con características tridimensionales y, por tanto, sólido o insoluble en todos los disolventes. Estos pueden hincharlo más o menos, según la capacidad de penetración. En la química en fase sólida se habla con frecuencia de la capacidad reactiva del soporte sólido, indicada generalmente por el número de miliequivalentes de funcionalización presentes en un gramo de soporte sólido. Por ejemplo, un poliestireno clorometilado de 1,2 miliequivalentes por gramo significa que un gramo de poliestireno contiene 42 miligramos de cloro reactivo (es decir, $3{,}38\times10^{-4}$ moles) y, por tanto, a cada gramo de polímero sólo podrán anclarse como máximo $3{,}38\times10^{-4}$ moles de un substrato, reactivo o catalizador en los puntos activos; el resto será inerte. Ello da idea de la débil capacidad reactiva, pero a la vez indica a qué separación se ven sometidas las moléculas ancladas en el polímero.

Son muchos los tipos de polímeros utilizados en la química en fase sólida. El más generalizado es el poliestireno tridimensional, insoluble, introducido más arriba. Otros polímeros de interés comercial utilizados son las poliamidas, la celulosa y los poliacrilatos.

El poseer un reactivo o un substrato anclado en un soporte insoluble confiere a la síntesis determinadas ventajas. En primer lugar, la separación del producto deseado se simplifica, ya que el soporte insoluble se elimina por simple filtración. Piénsese en el caso en que el polímero soporta la molécula deseada, separándose fácilmente el exceso de reactivos o los subproductos solubles; o aquel otro en que en el polímero quedan fijados los subproductos, obteniéndose la molécula deseada por filtración y evaporación subsiguiente del disolvente.

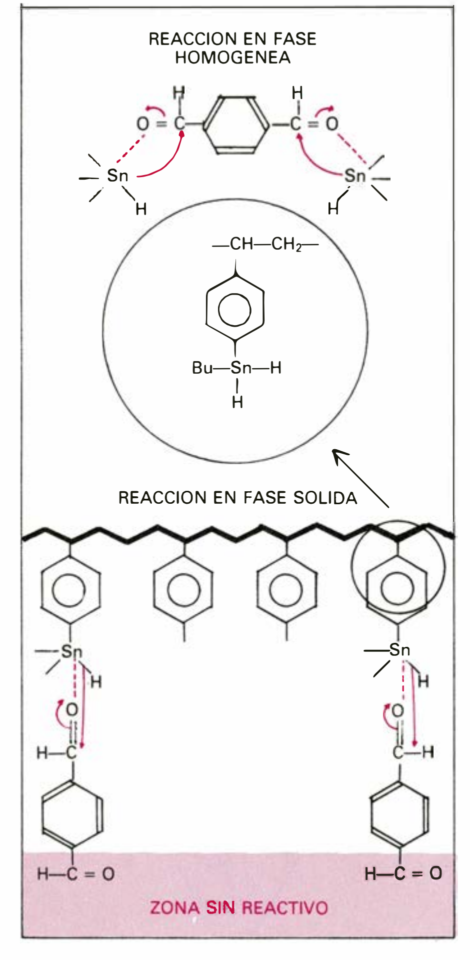
Otra ventaja es de orden económico: cuando se utiliza un reactivo costoso anclado sobre un polímero, el reactivo puede recuperarse para volver a emplearlo si no se consume (catalizadores). Asimismo, si la reacción procede hasta su totalidad, de modo que la única manipulación sea la filtración y lavado del polímero, la reacción puede automatizarse.

No acaban ahí las ventajas. Puesto que los polímeros entrecruzados son insolubles, y por tanto no volátiles e inodoros, su toxicidad disminuye considerablemente: ello significa que pueden hacerse ambientalmente aceptables aquellas reacciones en las que intervengan productos tóxicos o pestilentes. Además, los catalizadores soportados sobre polímeros permiten efectuar operaciones continuas.

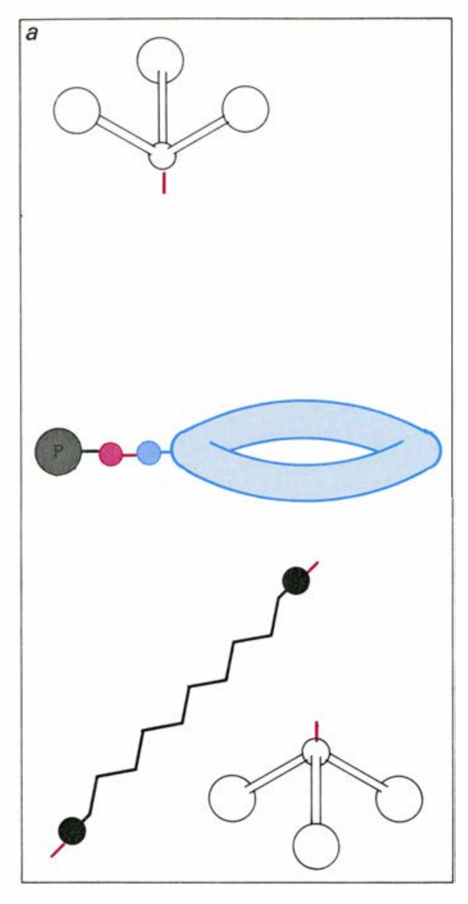
Por último, la fijación a un soporte polimérico de un reactivo o un substrato puede afectar —debido a condicionamientos estéricos y electrónicos de los microalrededores— al curso de la reacción química. Este punto, que conecta íntimamente con el especial modo de acción de los enzimas, no ha recibido todavía suficiente atención.

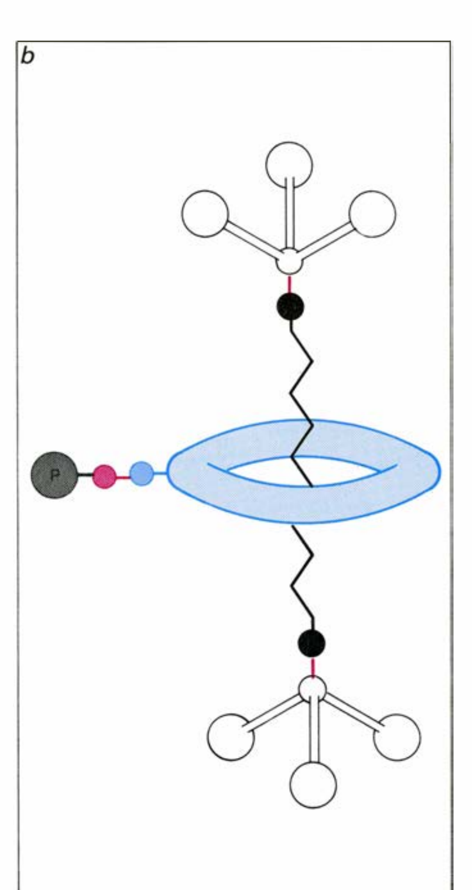
Claro está que existen también inconvenientes en el uso de la fase sólida en síntesis orgánica. Citemos la preparación del polímero debidamente funcionalizado o el adecuado anclaje de reactivos o substratos. Muchos centros reactivos se encuentran, además, en el interior de la trama polimérica, por lo que es necesaria la difusión de las moléculas reaccionantes hasta dichos centros; puesto que existen impedimentos estéricos y de solvatación o de humectación de la trama que dificultan esta aproximación, las reacciones por lo general son más lentas. Por otro lado, la reutilización del polímero no siempre reproduce los resultados de reaccionabilidad iniciales, ya que ocurren fragmentaciones y degradaciones de la propia trama polimérica.

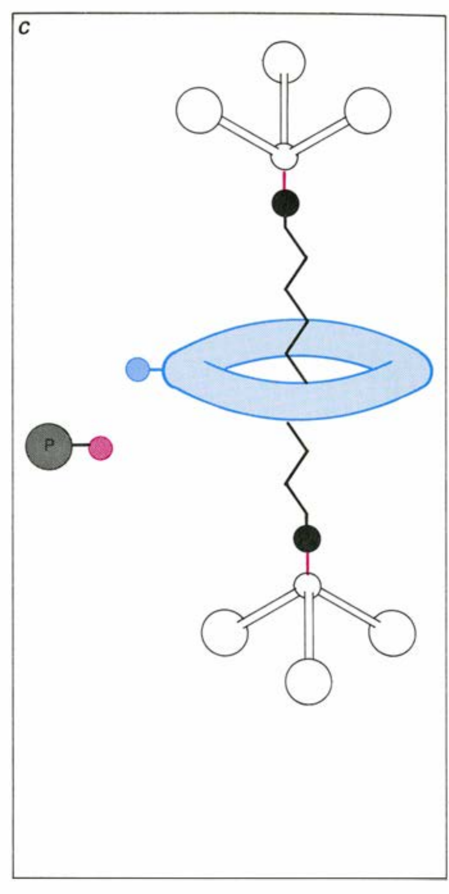
Otro inconveniente, muy serio, tiene que ver con la dificultad de llevar a cabo determinaciones analíticas: no es lo mismo efectuar análisis estructurales de una molécula libre (con la que pueden aplicarse todas las técnicas espectroscópicas y cromatográficas hoy en uso) que un análisis de una molécula anclada en un polímero. Su insolubilidad hace prácticamente imposible la utilización de las técnicas usuales de la resonancia magnética nuclear. No es posible efectuar separaciones cromatográficas. Sólo la espectroscopía infrarroja en fase sólida y el análisis elemental cuantitativo dan alguna información sobre el contenido de la molécula an-



6. REDUCCION DE TEREFTALDEHIDO (un dialdehído) con un hidruro de estaño. En fase homogénea, la reducción se efectúa sobre los dos grupos aldehídicos; en fase sólida, debido a los condicionamientos estéricos y a la congelación del substrato sobre el reactivo polimérico, puede lograrse una reducción selectiva de un solo grupo aldehído.







7. **"ENHEBRADO" DE MOLECULAS, conseguido con un soporte polimérico que llevaba anclado un anillo cicloalcánico de 30 átomos de carbono; al efectuar en presencia de este polímero la reacción de protección de los grupos alcohol de una molécula lineal difuncional (*a*), algunas moléculas penetran en el anillo y sufren el bloqueo de sus grupos funcionales, mientras permanecen en su interior; los grupos protectores, muy voluminosos, impiden que la molécula lineal escape, y queda enhebrada (*b*). Sin embargo, la probabilidad de que esto suceda es muy baja. No sería posible aislar moléculas enhebradas si la reacción se efectuara en fase homogénea, pues tales moléculas quedarían mezcladas en muy pequeña proporción con los dioles protegidos sin enhebrar. Pero efectuada la reacción tal como se indica en el esquema, todas las moléculas enhebradas quedan ancladas sobre el polímero; si la reacción se efectúa un número considerable de veces y se lava, de forma exhaustiva, el polímero de todos los reactivos iniciales, al someter a desanclaje el anillo cicloalcánico (*c*) aparece un tanto por ciento elevado de anillos enhebrados. Se ha logrado así una concentración de la operación de enhebrado sobre la fase sólida.**

clada. En algunos casos, debe efectuarse el desanclaje del reactivo o substrato y su posterior identificación en fase homogénea.

¿Qué tipos de reacciones se llevan a cabo con polímeros funcionalizados? Empezaremos por los reactivos. Los reactivos poliméricos se obtienen al anclar un reactivo (una sustancia que genera una reacción química concreta) sobre un polímero insoluble que actúa de soporte. Después de la reacción, el subproducto queda fijado en el polímero y puede filtrarse fácilmente. Los primeros reactivos anclados lo fueron de oxidación o reducción. Así, la epoxidación de olefinas (es decir, la adición de oxígeno a hidrocarburos insaturados) se realiza con perácidos orgánicos (perfórmico, *m*-cloroperbenzoico, etc.).

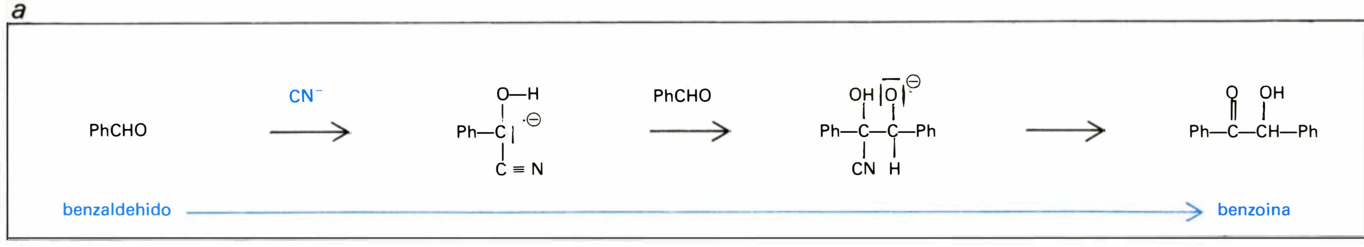
Así un ácido polimetacrílico, polímero que contiene grupos carboxilo (CO_2H) puede convertirse en perácido "polimérico" al tratarlo con agua oxigenada en presencia de un ácido sulfónico. El perácido, insoluble, al reaccionar con olefinas da epóxidos, pudiéndose recuperar, por simple filtración, el polímero que contiene de nuevo los grupos carboxilo no activos. Este soporte sólido puede regenerarse al ser tratado nuevamente con agua oxigenada. También pueden utilizarse funcionalizaciones de peroxidación ancladas a un soporte de poliestireno, es decir, P $-CO_3H$.

También se han descrito reactivos poliméricos reductores. Así, los hidruros de estaño son buenos agentes reductores de los grupos funcionales aldehídos y cetonas. Sin embargo, su uso viene limitado por su toxicidad y mal olor. Inconveniente que se salva usando un reactivo de estaño polimérico, que muestra la peculiaridad siguiente: cuando se le hace reaccionar con una molécula difuncional, por ejemplo tereftaldehído, sólo se reduce uno de los grupos aldehído, lo que se atribuye a que el substrato, en este caso el tereftaldehído, permanece unido al centro reactivo y a que la movilidad restringida del polímero, así como la lejanía de los grupos reactivos, no favorecen un doble ataque, fenómeno normal en soluciones homogéneas. Se logran de esta forma monorreacciones selectivas del grupo funcional. Diríase que aquí se está imitando a los enzimas.

A lo largo de los últimos años se han preparado muchos otros reactivos poliméricos. Los hay que halogenan (o introducen átomos de cloro o bromo) olefinas y cetonas, oxidan tioles y tioéteres, reducen disulfuros a tioles, convierten alcoholes en cloruros de alquilo y ácidos en cloruros de ácido, son utilizados para sintetizar glicósidos, inducen reacciones de condensación, etc., siempre soportados o anclados a una resina, generalmente poliestireno.

Considerados los reactivos, abordaremos los catalizadores poliméricos. Se contempla aquí la acción de catalizadores anclados en soportes sólidos. El catalizador no se consume a lo largo de la reacción; sólo participa rebajando la energía de activación y, por tanto, facilitando la velocidad de reacción sin modificar el equilibrio final.

a



b

Tiamina

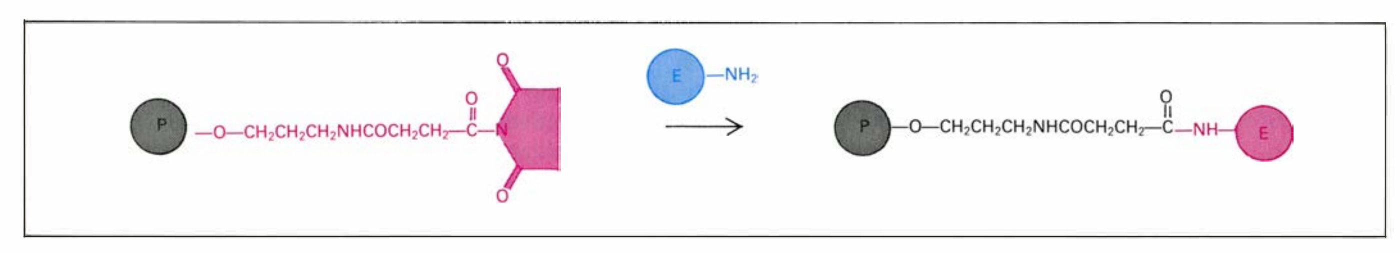
Base conjugada

c

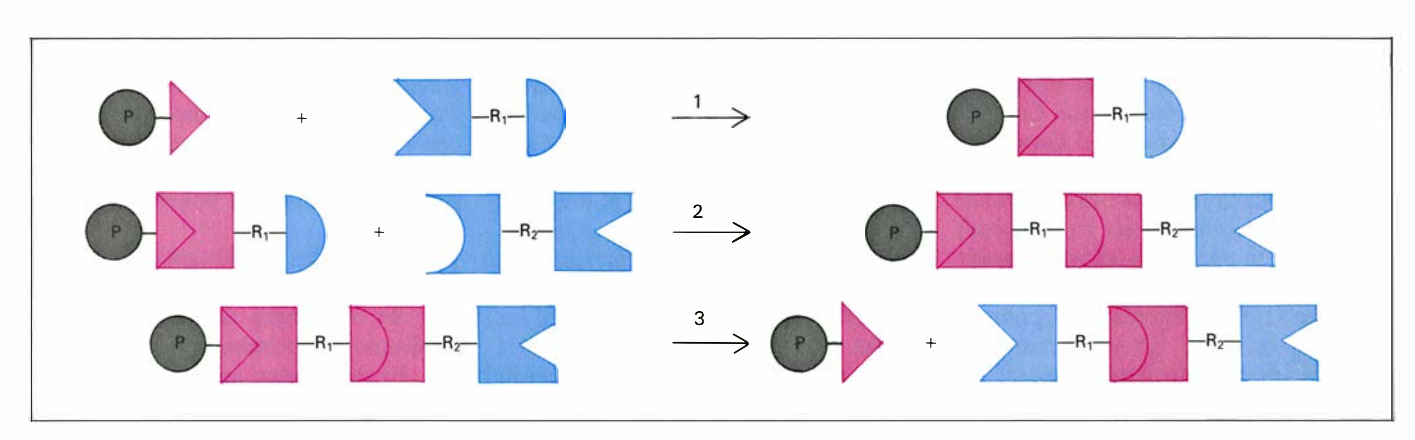
8. CONDENSACION BENZOINICA, reacción que se induce y cataliza mediante el ion cianuro sobre aldehídos aromáticos (*a*). Sin embargo, en los sistemas bioquímicos existe un coenzima, la vitamina B_1, o tiamina (*b*), que, en su forma de base conjugada, desempeña el mismo papel que el ion cianuro, catalizando condensaciones tipo benzoínico de aldehídos alifáticos, y haciéndolo de un modo estereoespecífico. La parte activa de la vitamina –la sal de tiazolio– ha podido anclarse sobre un polímero insoluble y se ha estudiado su eficacia como catalizador de las condensaciones de tipo benzónico.

Los primeros catalizadores poliméricos utilizados han sido resinas de intercambio iónico, tanto catiónicas como aniónicas, que poseen como base una matriz de poliestireno. Las resinas catiónicas, en realidad el ácido bencenosulfónico polimérico, pueden actuar de catalizador ácido en síntesis orgánicas donde se requiera la presencia de protones: reacciones de esterificación, de formación de acetales, de acilación y otras. La ventaja consiste en la fácil separación o eliminación del ácido del medio de reacción, a través de una simple filtración. Las resinas aniónicas, en su forma básica, han servido también de catalizadores.

Las reacciones catalizadas por estas resinas son a menudo mucho más limpias, esto es, proporcionan menos sustancias indeseables, que las correspondientes reacciones en fase homogénea. Se producen incluso fenómenos selectivos, como en la condensación catalizada por bases entre el formaldehído y aldehídos alifáticos: mientras en la condensación homogénea se obtiene una cantidad considerable de productos de autocondensación del aldehído alifático, el uso de una resina adecuada da lugar a una limpia condensación cruzada, entre el formol y el aldehído alifático. Una vez más, este comportamiento de las resinas mimetiza la acción de los enzimas. Y ello puede explicarse no sólo mediante consideraciones estéricas, o de anclaje del catalizador en una matriz sólida, sino también a través de fenómenos de adsorción o difusión de los



9. ANCLAJE O INMOVILIZACION de un enzima a un poliestireno convenientemente funcionalizado: el grupo terminal $-NH_2$ del enzima desplaza al grupo succinimida, con lo que el enzima se une covalentemente al polímero. De este modo, dicho enzima puede recuperarse con facilidad por filtración.



10. SINTESIS DE UNA MOLECULA ORGANICA sobre un soporte polimérico. El paso *1* (*primero, por arriba*) representa el anclaje de una molécula difuncional sobre el polímero, dejando libre una de las funciones a través de la cual se consigue concatenar un resto orgánico deseado (R_2, paso *2*). Este paso *2* puede unir nuevos restos o transformar los grupos funcionales. Finalmente, una vez se ha logrado la construcción de la molécula objetivo, en el paso *3* se indica la separación de esta molécula del polímero inicial. Con este procedimiento se ha sintetizado una variedad importante de moléculas.

distintos reactivos en la misma trama polimérica.

Ciertos catalizadores se unen, covalentemente, a los polímeros insolubles. Muestran una función distinta de la que cumplen los de la catálisis ácida-básica. Merecen destacarse los experimentos con sales de tiazolio que inducen la condensación benzoínica. Esta condensación se realiza normalmente en fase homogénea, por catálisis del ion cianuro sobre aldehídos aromáticos. En la naturaleza se realizan condensaciones de tipo benzoínico con aldehídos alifáticos, siendo el catalizador la base conjugada de la vitamina B_1, o tiamina, que es una sal de tiazolio específica; característica de estas condensaciones es su estereoespecificidad, es decir, el producto de condensación es quiral u ópticamente activo, en oposición a las reacciones convencionales realizadas en el laboratorio que conducen a racémicos o mezclas de los dos enantioisómeros. Esta estereoespecificidad es una peculiaridad importantísima de las reacciones enzimáticas.

En estrecha dependencia con los catalizadores están los fotosensibilizadores anclados en fase sólida. Son éstos ciertas sustancias orgánicas que absorben luz promocionándose a estados excitados, generalmente el triplete, desde donde pueden transferir energía electrónica a otras sustancias aceptoras que sufren entonces una reacción fotoquímica; sustancias aceptoras que no son capaces de absorber la luz utilizada en el experimento. Los fotosensibilizadores actúan, pues, de intermediarios en ciertas reacciones fotoquímicas y pueden considerarse como catalizadores de estas reacciones.

Uno de los fotosensibilizadores más utilizados, por ejemplo en reacciones de foto-oxidación, es el Rosa de Bengala, sustancia que se caracteriza por su complejidad estructural y su absorción de la luz visible. El inconveniente de su utilización es, una vez más, su separación de la mezcla de reacción final, perdiéndose. Su anclaje a un poliestireno entrecruzado clorometilado ha resuelto esta desventaja con un éxito tal que se ha comercializado ya su venta.

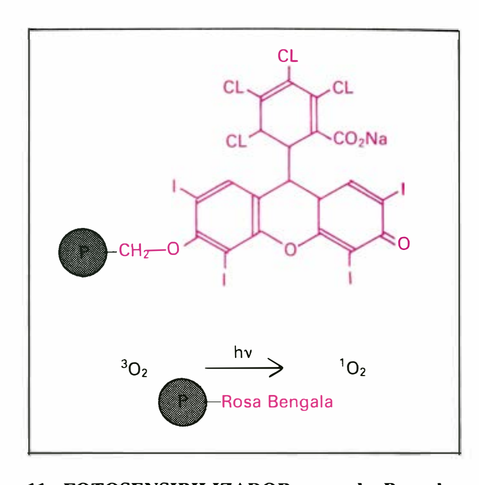
Puesto que los catalizadores por excelencia son los enzimas y normalmente estos compuestos son costosos, no es ilógico que se hayan anclado sobre soportes sólidos para evitar su pérdida y poder ser reutilizados. En el lenguaje bioquímico se habla entonces de enzimas inmovilizados. La inmovilización puede efectuarse por absorción del enzima en una fase sólida o por enlaces covalentes en polímeros hidrofílicos (es decir, en polímeros que se empapan en agua). Aunque esta técnica ya ha empezado a tener aplicaciones industriales queda mucho aún por hacer, sobre todo si se tiene en cuenta que el anclaje de un enzima ocasiona generalmente desnaturalizaciones que afectan a su poder catalítico. Por otro lado, la capacidad enzimática también viene afectada por los cambios en los microalrededores de los puntos activos, así como por las interacciones estéricas que puedan introducirse entre el enzima, el substrato y el soporte.

Abordados los reactivos y los catalizadores poliméricos, hora es de ocuparse de los substratos poliméricos. Siguiendo el modelo de la síntesis de péptidos nos vamos a referir en este apartado a aquellas reacciones de síntesis orgánica que se realizan de modo que el substrato se halla anclado sobre la fase sólida mientras los reactivos provienen de la solución homogénea que embebe o empapa al polímero. La molécula objetivo se construye o crece sobre el soporte insoluble.

Otro modo de visualizar este tipo de reacciones sería considerando el polímero como un grupo protector del substrato que va a reaccionar. Este substrato posee como mínimo una doble funcionalidad: por un lado, se utiliza para la unión con el polímero y, por otro, interviene en la reacción, a través de la cual se produce la verdadera síntesis. También aquí la principal ventaja de esta técnica es la facilidad de purificación de la fase sólida por lavado y filtración. Sin olvidar otras características, como la inmovilización o "congelación" de la molécula substrato sobre el soporte.

Los soportes poliméricos se escogen de acuerdo con su capacidad para intervenir en síntesis orgánicas. Las propiedades exigidas para tal menester se resumen en cinco: insolubilidad, para facilitar la purificación después de cada paso sintético; contener la funcionalidad accesible y reactiva que permita un fácil anclaje del substrato; el enlace entre el substrato y el polímero debe ser estable frente a los reactivos utilizados en la síntesis; dicho enlace podrá romperse con facilidad al final de la síntesis, sin que se afecten las otras funciones presentes en la molécula objetivo final, y, por último, el polímero se reutilizará después de conseguida la síntesis.

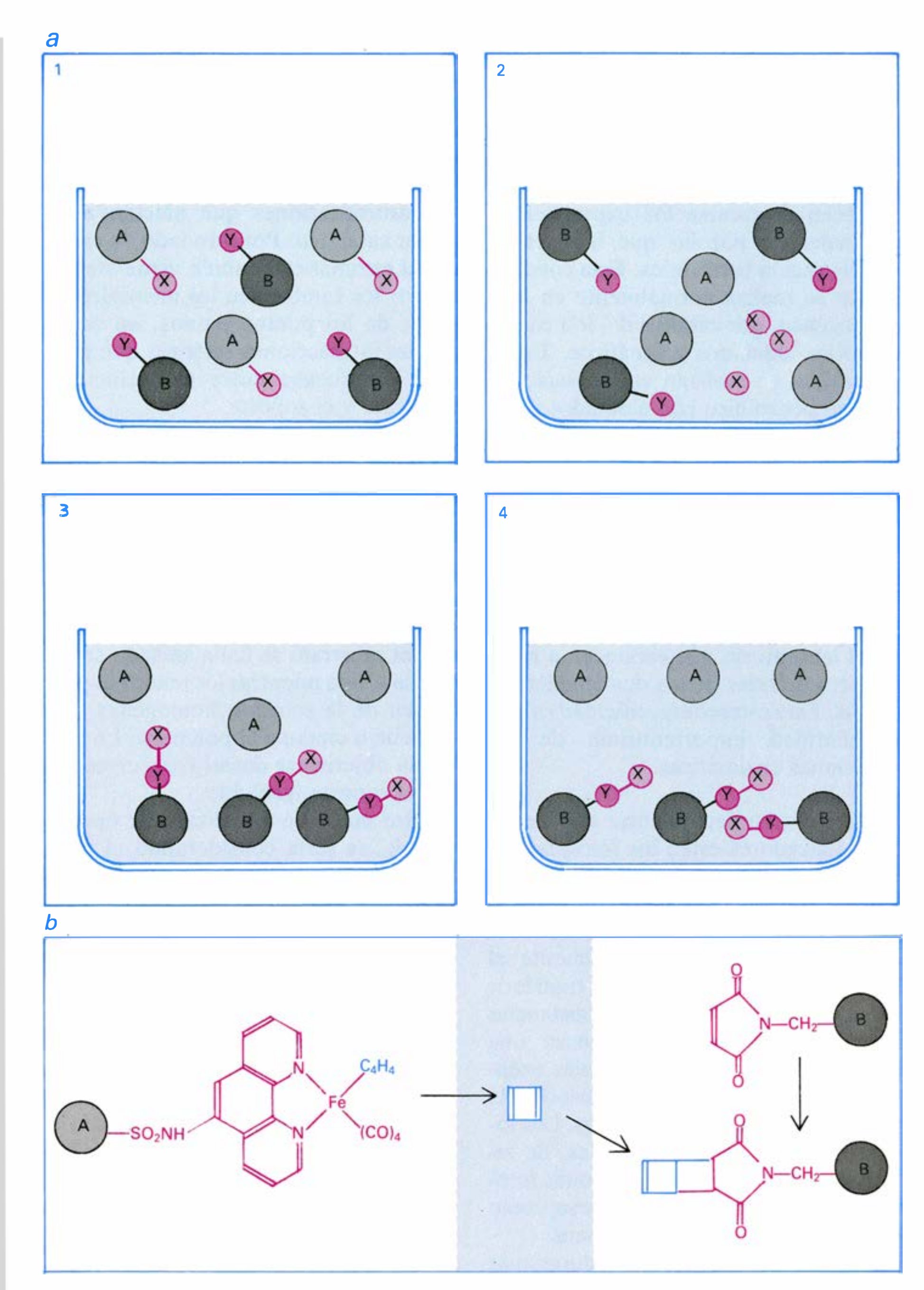
La química en fase sólida se ha venido aplicando a la síntesis de alcaloides, de azúcares, de carotenoides, de compuestos organometálicos, etcétera. Recordemos, a modo de ejemplo, lo conseguido en el campo de la síntesis de feromonas sexuales de insectos. Estas sustancias son emitidas específicamente por los insectos para atraerse y favorecer el apareamiento. Son sustancias volátiles, generalmente alcoholes grasos con una insaturación. El valor arquetípico del ejemplo reside en que en él se cumplen varios pasos sintéticos a través de los cuales se mantiene el substrato anclado sobre el polímero. El polímero, después de ulterior regeneración, puede volver a ser reutilizado. El esquema de dicha síntesis se basa en la monoprotección por parte del polímero de uno de los grupos hidroxilo del reactivo inicial, 1,8-octanodiol.



11. FOTOSENSIBILIZADOR rosa de Bengala. Bajo irradiación, convierte la molécula de oxígeno de estado fundamental triplete a oxígeno excitado, de estado singlete. El fotosensibilizador se ancla a un poliestireno entrecruzado e insoluble, para obtener una fácil recuperación del catalizador.

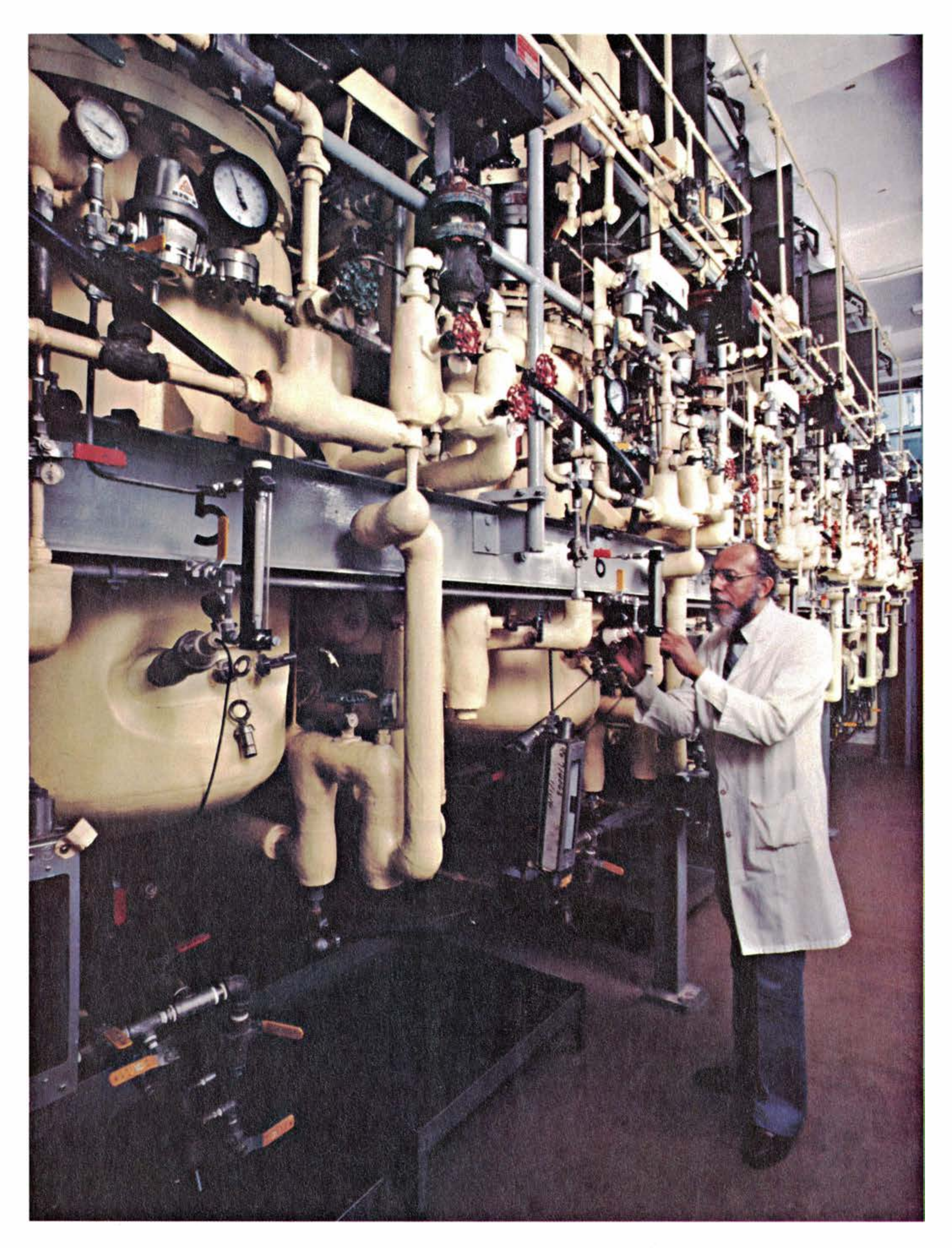
Los polímeros funcionalizados no sólo poseen aplicaciones en síntesis; se han empleado en estudios mecanísticos (es decir, en investigaciones dirigidas a determinar el curso de una reacción química) para probar la existencia de reactivos intermedios de vida muy corta, a través de una curiosa técnica que consiste en la utilización de dos polímeros funcionalizados de distinta porosidad y por tanto de distinta densidad, y en la suposición de que la reacción entre las funciones de dos perlas poliméricas es prácticamente imposible, dados los condicionamientos de distancias y contacto.

Sin embargo, si sobre el polímero *A* (de supuesta densidad mayor) se genera y libera un reactivo intermedio que pasa a la solución, el polímero *B* (de supuesta menor densidad) puede atraparlo siempre y cuando posea la funcionalización adecuada. La demostración de que tal ocurre nos ofrecería una prueba fehaciente de que el intermedio existe en libertad durante un lapso de tiempo, corto pero suficiente, y necesario para pasar de *A* a *B*. La distinta



12. TEST DE LAS TRES FASES para probar la existencia de intermedios reactivos. Se introducen en suspensión los dos polímeros (primer paso, *1*); se produce la reacción de liberación del intermedio, y éste aparece en el disolvente (*2*). El intermedio es captado por el substrato anclado en el segundo polímero (*3*). Al dejar en reposo la suspensión, los polímeros pueden separarse fácilmente y someterse a análisis (*4*).

densidad de *A* y *B* permitiría su fácil separación y el análisis independiente para probar que el intermedio ha sido captado por *B*. Hipótesis de trabajo que se ha visto ya confirmada con algunas sustancias; así, por ejemplo, el ciclobutadieno, molécula muy reactiva cuya vida media es de décimas de segundo. Aunque el químico está aún muy lejos de conseguir en su matraz las hazañas de los enzimas en el organismo, los procedimientos de la fase sólida le ofrecen unos métodos que mimetizan en cierto modo la acción de los catalizadores bioquímicos. En la década de los setenta se exploró y aplicó el método a reactivos y a determinados pasos de síntesis, ya que su principal ventaja reside en la fácil separación del polímero insoluble, evitando tediosos procedimientos de separación, pérdidas de reactivos caros y contaminaciones ambientales. Si bien no se ha encontrado una aplicación industrial generalizada, los próximos años pueden ser testigos de ello, como épocas no lejanas lo fueron de las resinas intercambiadoras de iones, parientes cercanos de los polímeros usualmente utilizados en la química en fase sólida.



1. INTERFERON FABRICADO por bacterias en una batería de tanques de fermentación de Hoffmann-La Roche, Inc., de Nutley, Nueva Jersey. Se desarrollan en un medio de cultivo células de *Escherichia coli* en las que se ha insertado ADN recombinante que transporta un gen para el interferón humano alfa. Las bacterias proliferan, sintetizando interferón junto con miles de sus proteínas propias. Cuando las células bacterianas alcanzan una concentración máxima, se sacrifican, se sacan de los tanques de fermentación y se concentran por centrifugación; se cobra luego el interferón sintetizado.

Purificación y fabricación de interferones humanos

No han satisfecho todas las esperanzas depositadas en ellos, pero sus genes ya se han aislado y clonado en E. coli. *Las bacterias producen abundante interferón, sometido ahora a ensayos clínicos en humanos*

Sidney Pestka

En la década de 1930, varios investigadores definieron el fenómeno de interferencia vírica: la infección que, en un animal, provocaba un virus parecía protegerle frente a una posterior infección por otro virus. En 1957, Alick Isaacs y Jean Lindenmann, del Instituto Nacional de Investigaciones Médicas de Londres, encontraron un agente de interferencia vírica, una proteína liberada por células expuestas a un virus, que permitía a otras células adquirir resistencia a la infección vírica. A esa sustancia la llamaron interferón.

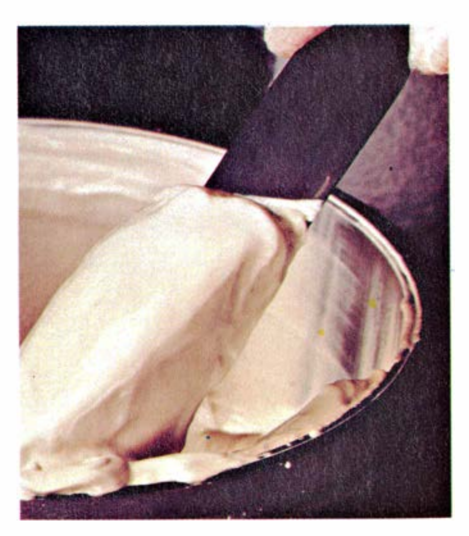
Las posibilidades antivíricas del interferón quedaron de manifiesto desde el mismo instante de su hallazgo. Lo que se corroboró en otra publicación donde se informaba de logros similares realizados por Y. Nagano y Y. Kojima, del Instituto de Enfermedades Infecciosas de Tokio. Las esperanzas se cifraban en que la proteína no actuaba contra ningún virus; antes bien, protegía a la célula frente a una amplia gama de éstos. Además, afectaba a varias funciones celulares por caminos que sugerían distintas posibilidades terapéuticas. Se trataba de una sustancia potente: una pequeñísima cantidad de ella ejercía una acción prolongada. Al ser un producto natural se le juzgó (administrado a dosis correctas) más inocuo que la mayoría de las nuevas drogas experimentales.

Mas aquella euforia inicial se ha enfriado con el paso del tiempo y la falta de resultados. Veinte años después de su descubrimiento, siguen apareciendo nuevas dificultades y problemas complejos, junto con perspectivas optimistas. Se vio, por un lado, que el interferón no era una sola proteína, sino una gran familia de ellas, que variaban de una especie a otra y que, incluso dentro de una misma especie, se presentaban de múltiples formas. Su modo de acción, que a duras penas conocemos, resultó ser indirecto. Y, sobre todo, las células lo segregaban en cantidades diminutas, lo que convertía su purificación en tarea sumamente difícil. Durante muchos años pareció imposible lograr suficiente cantidad de interferón para hacer pruebas clínicas eficaces. Lo más que se conseguía eran extractos crudos que contenían varios interferones y gran cantidad de otras proteínas. Sin purificación no había forma de establecer la actividad propia de cada interferón y determinar su estructura, lo que nos hubiera permitido clasificarlos y asegurar su inocuidad y eficacia.

El panorama ha cambiado a lo largo de los últimos años. Se purificaron varios interferones humanos; y con ello se nos ha dado conocer su estructura y clasificar y valorar sus actividades. Al poco llegó la posibilidad de utilizar técnicas de ADN recombinante para aislar genes que determinan el interferón humano, clonarlos en bacterias, producir en cantidad notable interferón humano recombinante, por fermentación, y purificarlo por medio de anticuerpos monoclonales. En 1981 comenzaron los primeros ensayos para establecer las dosis adecuadas y observar los efectos secundarios del producto bacteriano purificado en humanos. Hoy hay ya en marcha ensayos de vasto alcance para comprobar la eficacia del interferón frente a enfermedades producidas por virus y frente a diversos tipos de cáncer. Se ha producido una rápida transferencia de la nueva tecnología biológica del laboratorio al ensayo farmacéutico.

Isaacs y Lindenmann descubrieron el interferón producido por células de pollo expuestas al virus de la gripe. Vieron que protegía a las células de pollo, pero no a las células de otros animales, frente a la infección vírica. Aunque el interferón no se mostraba específico para el virus (en consonancia con el concepto de interferencia vírica), sí ofrecía la llamada especificidad de especie; es decir, cada especie animal producía su propio interferón. Las cosas se complicarían más tarde al encontrarse que cada especie fabricaba varias clases de interferón, cada una de ellas con su espectro particular de actividad en otras especies.

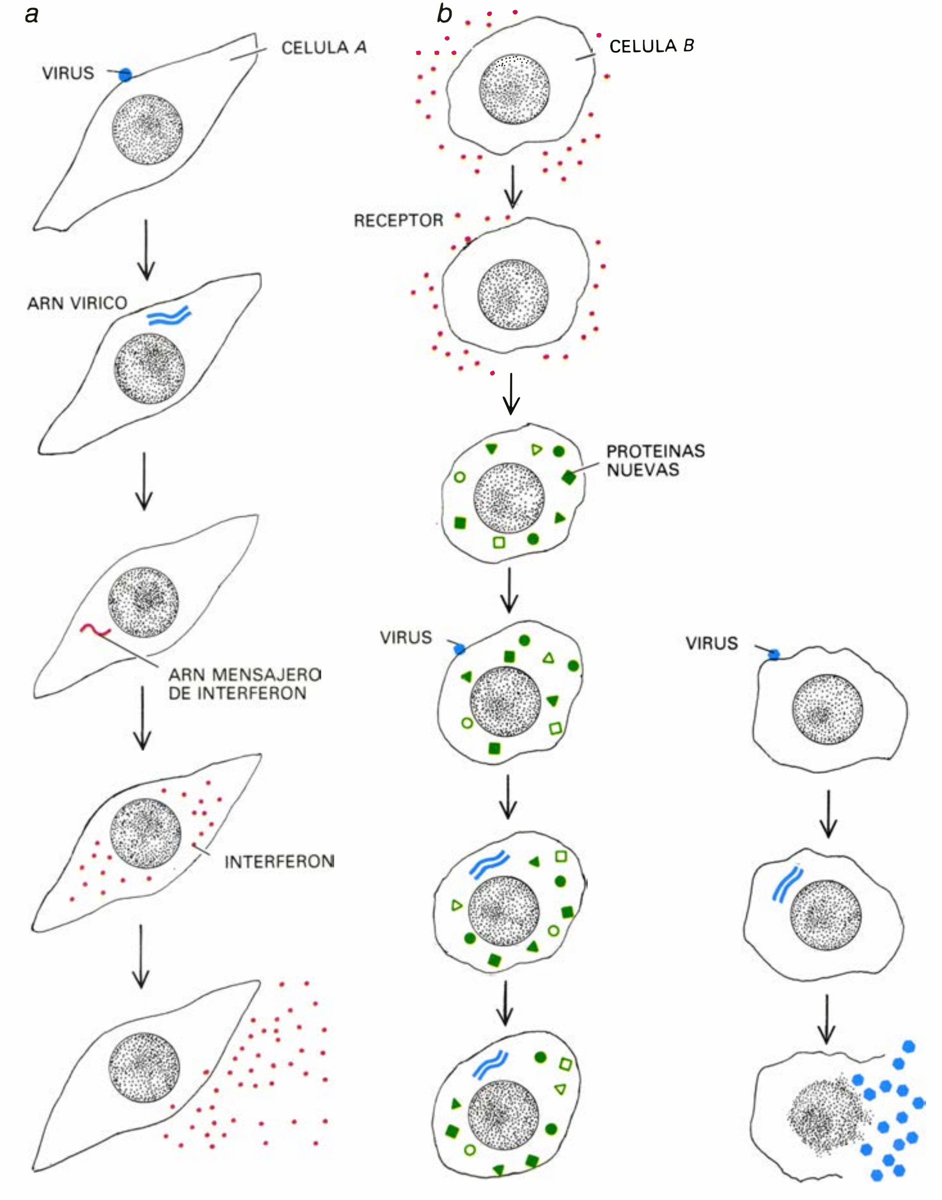
Me ceñiré aquí a los interferones humanos, divididos hoy nítidamente en tres clases. Por razones históricas, se les conoce desde hace tiempo por interferón de leucocito, interferón de fibroblasto e interferón inmune. Sabemos ya que las distintas clases no se producen en tipos celulares específicos, por lo que se ha propuesto un nuevo tipo de nomenclatura: alfa (para el interferón



2. PASTA DE *E. COLI* sacándose de la vasija de centrifugación. Las células se resuspenden y se rompen, tras lo cual los restos celulares se separan por centrifugación. Los ácidos nucleicos y otros materiales viscosos se eliminan; las proteínas del extracto pasan a través de una columna con anticuerpo monoclonal, para purificar el interferón.

de leucocito), beta (para el de fibroblasto) y gamma (para el interferón inmune), lo que representa todavía una simplificación forzada, pues se han aislado y clasificado muchos miembros distintos de la clase alfa. Algunos trabajos pretenden hallar subdivisiones en la clase beta. Hasta la fecha, sólo se ha aislado y purificado un miembro de la clase gamma.

La razón de que el efecto protector del interferón no se limite a un solo virus estriba en que (a diferencia de un anticuerpo) no interactúa con el virus atacante; el interferón interactúa con la célula a la que protege. La síntesis de esta proteína se induce cuando un virus introduce en la célula su material genético (ADN o ARN). La presencia del material extraño (probablemente en forma de ARN bicatenario) induce a la célula a sintetizar y segregar moléculas de interferón. El interferón segregado se une a un receptor específico de la superficie de otras células y, con ello, parece desencadenar varios mecanismos en las células receptoras, que inician la síntesis de distintas proteínas (algunas de las cuales se han identificado). Las proteínas confieren a las células resistencia contra los efectos habituales de la infección vírica: multiplicación del virus, lisis (estallido) de la célula y liberación de la descendencia de virus. Es mucho lo que queda todavía por conocer acerca del modo de acción del interferón, pero los experimentos de Thomas C. Merigan Jr., de la Facultad de Medicina de la Universidad de Stanford, y de otros han dejado claro que incluso los extractos crudos ejercen efectos clínicos frente a algunas infecciones víricas.

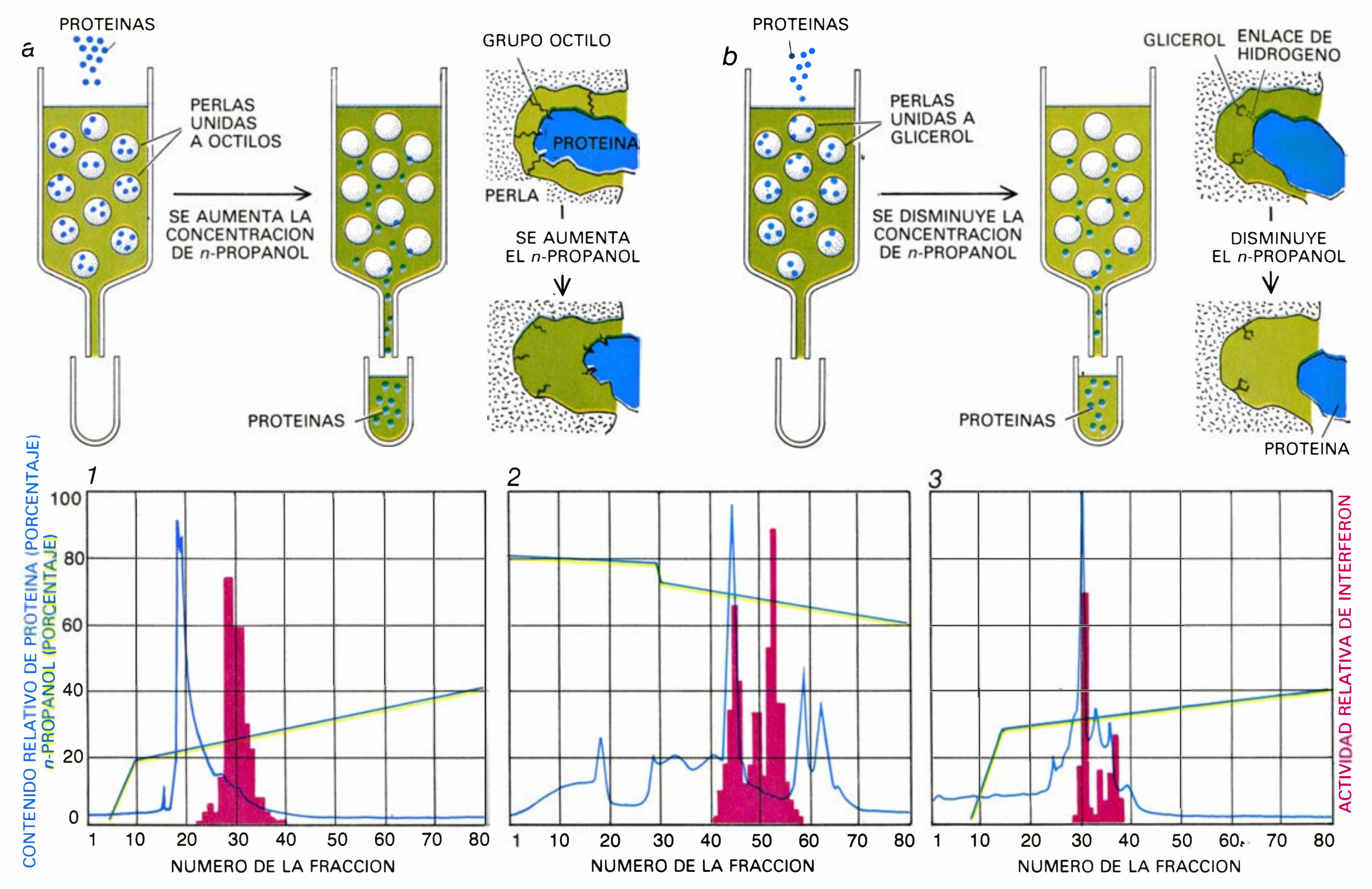


3. ETAPAS DE LA SINTESIS Y ACTIVIDAD ANTIVIRICA del interferón. La proteína se fabrica y secreta (*a*) en una célula infectada por un virus. Por lo que parece, la presencia de ARN vírico bicatenario en la célula induce la síntesis de ARN mensajero de interferón, que luego se traduce a interferón. El interferón segregado no ataca directamente al virus, sino que se une a moléculas receptoras específicas instaladas en la superficie de otra célula (*b*). La unión del interferón desencadena numerosos cambios en la actividad celular, incluida la síntesis de proteínas que confieren a la célula resistencia frente a la infección por virus. En la célula no protegida (*derecha*), el virus se replicaría haciéndola estallar.

Como he indicado antes, el interferón posee múltiples actividades biológicas, amén de las que contribuyen al efecto antivírico. Dos de ellas revisten particular interés. La proteína parece inhibir, en muchos casos, la proliferación de las células y estimula, además, la actividad de ciertas células del sistema inmunitario de los mamíferos: los linfocitos, denominados asesinos naturales de células, que actúan en la destrucción de células extrañas y, tal vez, de células cancerosas. Cualquiera de estos efectos, o ambos, podrían explicar los resultados obtenidos a lo largo de los años, especialmente por Hans Strander, del Instituto Carolino de Estocolmo, según los cuales el interferón promueve la regresión de algunos tumores.

Estudios esporádicos sobre el cáncer y los más directos que tratan de asegurar la eficacia del interferón frente a las enfermedades víricas se resintieron de la escasez y el elevado coste del interferón y, lo que era peor, de que el "interferón" disponible constituía una mezcla de varias proteínas, en la que menos del 1 por ciento en peso era interferón propiamente dicho. Aparte de la actividad antivírica, ningún otro efecto de la mezcla podía atribuirse con seguridad al interferón. Por ello, la purificación de la proteína aparecía como una cuestión de la máxima prioridad.

En 1977 emprendimos la tarea de purificar interferón alfa humano en mi laboratorio del Instituto Roche de Biología Molecular. Había que comenzar por asegurarse abundante cantidad de interferón de leucocitos en bruto. Nuestro método de preparación no difería, en grandes líneas, del desarrollado varios años antes por Kari Cantell, del Centro Hematológico de la Cruz Roja finlandesa. Los leucocitos (glóbulos blancos sanguíneos) humanos se incuban con un virus inductor, bien sea el virus Sendai o el virus de la enfermedad de Newcastle. Sustituimos el suero sanguíneo humano o bovino, utilizado por Cantell como componente del medio de cultivo, por la proteína láctica caseína; al ser ésta una sola proteína, resulta más fácil de eliminar, en los pasos iniciales de concentración, que las diversas proteínas del suero. Vimos que podía aumentarse el rendimiento susti-



4. PURIFICACION DEL INTERFERON por medio de cromatografía líquida de alta eficacia (*arriba*). En el método de la fase inversa (*a*) se empaquetan en una columna de cromatografía perlas de sílice a las que se han unido grupos octilo. Las proteínas que se hacen pasar a través de la columna se unen fuertemente a los grupos octilo. Descienden por la columna concentraciones crecientes del solvente *n*-propanol en solución acuosa; el *n*-propanol libera las proteínas de la columna, siguiendo el orden de su creciente afinidad por los grupos octilo. En el método de fase normal (*b*) se une glicerol a las perlas. En presencia de una elevada concentración de *n*-propanol, las proteínas que atraviesan la columna forman enlaces de hidrógeno con el glicerol. Los datos del proceso (*abajo*) muestran la alternancia de cromatografía de fase inversa (*1*, *3*) con la cromatografía de fase normal (*2*). Se controlaba el contenido proteico de las sucesivas fracciones, y se ensayaba en ellas la actividad de interferón (*barras rojas*). Finalmente se aisló una fracción de interferón puro (*3*).

tuyendo los leucocitos normales por leucocitos de pacientes afectos de leucemia mielógena crónica. Obtuvimos gran cantidad de células leucémicas (que se eliminan durante la terapia de la enfermedad) del Hospital M. D. Anderson y del Tumor Institute de Houston.

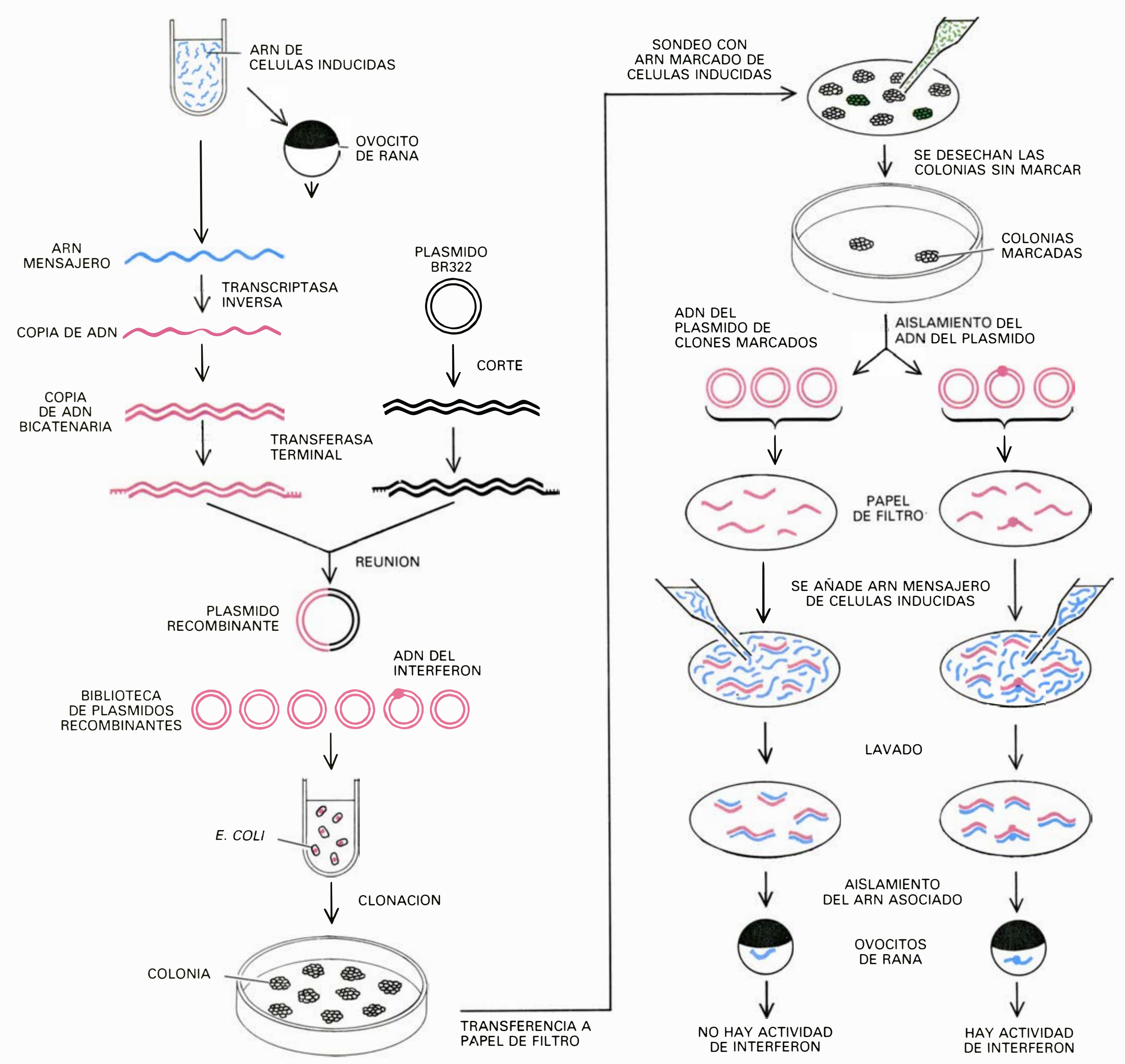
Tras una noche de incubación, las células y los virus se eliminan por centrifugación. Queda entonces una preparación cruda de interferón que contiene algo de interferón inducido junto con las demás proteínas cuya síntesis haya podido inducir el virus, además de todas las secreciones de la célula normal. El objetivo se centra entonces en eliminar, a lo largo de una serie de pasos de purificación, todas las proteínas que no sean interferón. En cada paso es preciso valorar, en el concentrado, la actividad del interferón en términos de unidades convencionales de "actividad específica". El ensayo habitual, en el que se mide la cuantía en que una muestra inhibe la destrucción de células por virus, duraba tres días. Philip C. Familletti y Sara Rubinstein encontraron un medio para acortar el tiempo hasta menos de 16 horas, con lo que se aligeró considerablemente el proceso de purificación.

Conscientes de las grandes dificultades que entrañaba purificar interferón a través de las técnicas convencionales, decidimos probar lo que se denomina cromatografía líquida de alta eficacia. Los métodos cromatográficos implican, en general, la adsorción de un extracto crudo sobre un soporte sólido y la elución de las diferentes fracciones con un solvente. En la cromatografía líquida de alta eficacia, la mezcla de partida se adsorbe sobre partículas, o perlas de vidrio, muy finas, empaquetadas en una columna, y se hace pasar el solvente a través de la columna. Sidney Udenfriend, Stanley Stein y Peter Böhlen, de Roche, habían separado péptidos (cadenas cortas de aminoácidos) por medio de cromatografía líquida de fase inversa, en la que la fase sólida (las perlas de vidrio) se recubrían con un material orgánico hidrófobo, que repele el agua; la fase móvil (el solvente) era más polar, es decir, hidrófila. Stein, Menachem Rubinstein y el autor decidieron aplicar el método para purificar el interferón alfa.

El alcohol etílico, solvente elegido, no funcionó; la proteína quedaba adsorbida sobre las perlas de vidrio. Decidimos probar un solvente algo menos polar, el *n*-propanol, aun cuando las proteínas no eran muy solubles en él, y podía darse el caso de que el interferón precipitara al ir eluyéndose. El gradiente de *n*-butanol funcionó. A medida que bombeábamos concentraciones crecientes del solvente a través de la columna, las diferentes fracciones proteicas (con diferentes afinidades por el soporte sólido) se eluían sucesivamente y se recogían en tubos. En cada fracción se valoraba la actividad del interferón. Las fracciones relativamente ricas en interferón se aplicaban a otra columna para posterior purificación. Alternando el proceso de fase inversa con la cromatografía de fase normal (en la que las perlas de vidrio están recubiertas con grupos hidrófilos y el solvente es menos polar), Rubinstein purificó interferón

alfa humano, en sólo unos pocos pasos, unas 80.000 veces. La actividad específica del interferón purificado rendía entre 200 y 400 millones de unidades por miligramo. Sometido a electroforesis en gel, que separa las proteínas en razón del tamaño molecular, el interferón dio una sola banda, correspondiente a un peso molecular de 17.500; la proteína de esa banda única era activa. Dicho de otra manera, habíamos purificado interferón alfa humano hasta la homogeneidad.

Durante el proceso de purificación nos había sorprendido observar que la actividad del interferón se purificaba en varias fracciones diferentes, señal de que estábamos aislando varias especies diferentes de la proteína. La caracterización química de los productos purificados lo corroboró. Al tratar la proteína con tripsina, un enzima que corta la larga cadena de subunidades de aminoácidos en sitios específicos, los diferentes productos purificados produjeron conjuntos de péptidos algo diferentes. Cuando dispusimos de cantidad suficiente de los distintos interferones purificados, procedimos a analizar



5. CLONACION DEL GEN DEL INTERFERON, que se inicia con la inducción de la síntesis de interferón en leucocitos. Se ensaya, en ovocitos, la actividad de interferón del ARN mensajero de las células. Si en una muestra aparece actividad de interferón, se obtiene la copia del ARN en ADN y luego se fabrica una segunda cadena, con ADN polimerasa. El "ADN copia" (ADNc) se trata con transferasa terminal para producir en ella "extremos cohesivos". Simultáneamente, el plásmido BR322, utilizado como vector, se corta, se abre y se sintetizan en él extremos cohesivos complementarios. El ADNc se hibrida con el ADN del plásmido para formar una "biblioteca" de plásmidos recombinantes, que luego se introducen en células de *E. coli*, que producirán clones de células idénticas. Para encontrar un clon que albergue plásmidos con el ADN correspondiente al interferón, las colonias, replicadas sobre papel de filtro, se bañan en ARN marcado radiactivamente procedente de células inducidas. El ARN sonda se hibridará, identificándolo, con el ADN que codifique proteínas inducidas, interferón incluido. Se corta en fragmentos lineales el ADN del plásmido de las colonias marcadas, que se fijan, en lotes, sobre papel de filtro. Se añade entonces ARN de células inducidas, que se hibridará con el ADN que codifica las proteínas inducidas. Se separa del ADN el ARN hibridado y se inyecta en ovocitos para identificar los lotes de ADN de plásmido que codifican interferón. El proceso se repite para cada clon de los lotes positivos.

su composición de aminoácidos (el número de cada uno de los 20 aminoácidos) mediante un analizador de aminoácidos muy sensible que había construido Stein. De nuevo se comprobó la presencia de varias especies.

Informes anteriores hablaban de preparaciones crudas de interferón alfa humano con ciertas diferencias en cuanto a la carga eléctrica de las moléculas, pero no se relacionaron con variaciones en la secuencia de aminoácidos de la molécula, que es el determinante último de la individualidad de una proteína. Se pensaba que los interferones eran glicoproteínas (proteínas a las que va unida una cadena azucarada), y se creía que la heterogeneidad de la carga respondía a diferencias en el contenido de carbohidrato. John A. Moschera examinó cinco especies alfa purificadas y no encontró carbohidrato detectable en ninguna de ellas. Al poco, el grupo de Christian B. Anfinsen, del Instituto Nacional de Artritis, Metabolismo y Enfermedades digestivas de los Estados Unidos, y Geoffrey Allen y Karl H. Fantes, de los laboratorios Wellcome de Gran Bretaña, aislaron múltiples especies de interferón alfa. Allen y Fantes buscaron en vano la presencia de carbohidrato. Ciertamente, algunas especies de interferón pueden estar glicosiladas, pero no es válido afirmar que todos son glicoproteínas.

Conocida la composición de aminoácidos, varios grupos se empeñaron en establecer su secuencia. Kathryn C. Zoon y sus colegas, del laboratorio de Anfinsen, colaborando con Michael W. Hunkapiller y Leroy E. Hood, del Instituto de Tecnología de California, determinaron la primera secuencia parcial del extremo amínico (el comienzo) de su interferón alfa humano. Pocos meses después, Warren P. Levy, de mi laboratorio, y John E. Shively, del City of Hope Medical Hospital, determinaron la secuencia amino-terminal de una de nuestras especies alfa. En dos sitios, la secuencia de Zoon y la nuestra mostraban un aminoácido diferente. Los aminoácidos implicados no sugerían la posibilidad de una confusión; por lo que admitimos que ambas secuencias eran correctas. Se disponía así de otra prueba de que los interferones alfa conformaban una familia de proteínas estrechamente relacionadas, hallazgo que se confirmó cuando Levy y Shively, Zoon y sus colegas, y Allen y Fantes obtuvieron más secuencias.

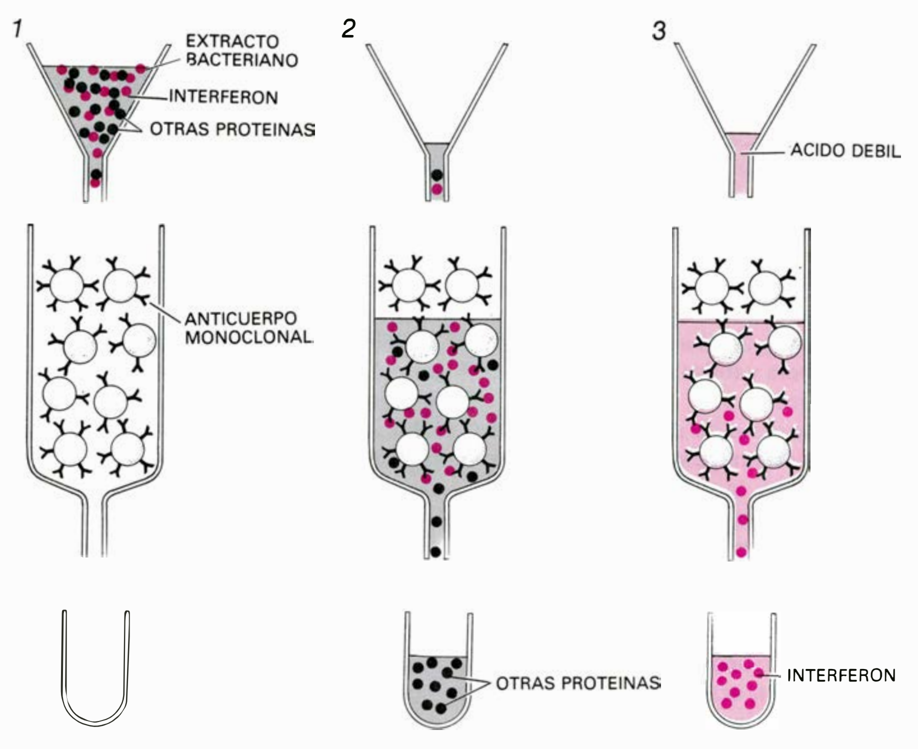
Aunque nuestro interferón alfa fue el primero cuyo grado de purificación permitió determinar su composición, Ernest Knight Jr., de la E.I. du Pont de Nemours & Company, había purificado algo de interferón beta humano (de fibroblasto) unos años antes. Knight, Yin H. Tan, de la Universidad de Calgary, y Stein, Heinz-Jürgen Friesen y Moschera determinaron, en mi laboratorio, las secuencias terminales del interferón beta purificado. Todas las secuencias resultaron iguales; por el momento cada grupo ha aislado un solo interferón beta y todas las proteínas parecen idénticas.

Al disponerse de interferón humano alfa y beta homogéneos, pudo establecerse, por fin, la actividad biológica de los interferones puros; y dejar de lado la de mezclas crudas. Se comprobó que la actividad antivírica de los interferones no era estrictamente específica de la especie. La mayoría de los interferones alfa (de leucocitos) humanos resultaron casi tan activos en la protección de células bovinas como en la de células humanas; de hecho, dos de ellos se evidenciaban mucho más activos sobre células bovinas que sobre células humanas. Sin embargo, el interferón beta (de fibroblastos) humano presentaba una actividd antivírica muy baja sobre células bovinas. Trabajando con nuestras especies alfa purificadas, Marian Evinger, en mi laboratorio, y Nathan O. Kaplan, de la Universidad de California en San Diego, encontraron que el interferón presentaba también actividad antiproliferativa, pero que, en ese sentido, unas versiones de la proteína eran mucho más eficaces que otras. Al igual que en la estimulación de la actividad para matar células, Ronald B. Heberman y John R. Ortaldo, del Instituto Nacional del Cáncer, encontraron que casi todos los interferones puros exaltaban la capacidad de ciertos linfocitos para destruir células diana.

Quedaba por resolver el problema de la obtención de interferón puro en cantidad suficiente para abordar su modo de acción y realizar ensayos clínicos. Podíamos haber recogido el necesario a través de la inducción a gran escala de leucocitos y purificación de material crudo; pero disponíamos de un método mejor. Hacia mediados de la década de 1970, el desarrollo de las técnicas del ADN recombinante permitía ya insertar en la bacteria *Escherichia coli* el gen de una determinada proteína, para clonarlo, o replicarlo en muchas copias, y expresarlo luego: traducirlo a proteína. La técnica nos permitiría obtener gran cantidad de genes de un determinado interferón, para su estudio, y obligar a la bacteria que fabricase un interferón determinado, quizá, a la larga, a escala industrial.

La verdad es que habíamos dado un primer paso varios años antes. Suele empezarse el proceso de clonación de un gen aislando, de células que fabrican gran cantidad de la proteína deseada, el ARN mensajero que codifica esa proteína. (El ARN mensajero es el ácido nucleico fruto de la transcripción celular del ADN de un gen; posteriormente la maquinaria celular lo traduce a proteína.) El ARN mensajero del interferón se presenta en los leucocitos en cantidades diminutas; de ahí que cueste identificarlo. En 1974, James L. McInnes, de mi laboratorio, trabajando con Edward A. Havell y Jan Vilček, de la Facultad de Medicina de la Universidad de Nueva York, extrajeron ARN mensajero de células obligadas a producir interferón. Añadieron el ARN a extractos libres de células, que contenían la maquinaria celular para la traducción, y lograron traducir el ARN y obtener interferón activo. Más tarde, Ralph L. Cavalieri insertaba el ARN mensajero en ovocitos (precursores de los óvulos) de rana; los ovocitos traducían el mensajero. Ensayando el producto del sistema libre de células, o de los ovocitos, en cuanto a su actividad interferoniana, determinábamos si una preparación de ARN contenía al menos algo de mensajero de interferón. M. N. Thang, del Instituto de Biología Fisicoquímica, y Paula M. Pitha, de la Facultad de Medicina de la Universidad Johns Hopkins, desarrollaron métodos similares para valorar el ARN mensajero del interferón.

En cuanto Familletti optimizó la producción de ARN mensajero del interferón en leucocitos inducidos, Shuichiro Maeda, Russell McCandliss y Alan Sloma prepararon ADN complementario a partir del ARN e insertaron el ADN en plásmidos, pequeños anillos de ADN bacteriano. Se reunió así una gran "biblioteca" de plásmidos recombinantes, cada uno de los cuales llevaba una copia de ADN (ADNc) de alguno de los ARN mensajeros procedentes de los leucocitos inducidos. Se introdujeron los plásmidos en células de *E. coli* y se aislaron los clones resultantes (colonias descendientes de una sola célula). El problema siguiente –lo más duro del procedimiento– era encontrar, en la amplia biblioteca de plásmidos recombinantes, aquellos cuyo ADN codificaba el interferón.



6. ANTICUERPOS MONOCLONALES utilizados para purificar interferón recombinante. Se vierte en una columna repleta de perlas de vidrio, a las que previamente se han unido anticuerpos monoclonales específicos contra interferón alfa, una mezcla de interferón crudo extraída de las bacterias (*1*). El interferón se une a los anticuerpos y permanece ligado mientras las demás proteínas atraviesan la columna (*2*). Un ácido débil desaloja el interferón (*3*); la solución se neutraliza y se procede a concentrar el interferón.

Si hubiésemos partido de un ARN mensajero del interferón ya purificado, la tarea hubiera sido mucho más sencilla. El ADN y el ARN son cadenas de ciertas subunidades denominadas nucleótidos; la secuencia de los nucleótidos codifica la información genética. El ARN transcrito a partir de un fragmento de ADN es complementario de éste en la secuencia de nucleótidos, y se hibridará con él, apareándose las dos cadenas para formar una doble cadena. Un ARN de interferón puro puede, por tanto, marcarse con un isótopo radiactivo y utilizarse como sonda que se reasocia en torno al ADN que se busca, cuya identidad queda revelada por autorradiografía.

El lector recordará, sin embargo, que no poseíamos ARN de interferón puro; habíamos empezado con una mezcla de distintos ARN cuya actividad interferoniana específica sólo podía ponerse de manifiesto después de traducir el ARN y calibrando, en los productos, la actividad de interferón. Había, por tanto, que diseñar un procedimiento indirecto, en dos pasos. En el primer paso se revisaban todas las colonias bacterianas, para averiguar cuáles portaban plásmidos con ADNc formado a partir del ARN de células inducidas y que, por tanto, quizá transportaran ADNc que codificara interferón. Efectuamos el ensayo usando, por sonda, ARN marcado de células inducidas, en presencia de gran cantidad de ARN sin marcar, procedente de células no inducidas, de suerte que la sonda marcada se hibridase solamente con el ADNc inducido. La búsqueda nos permitió desechar casi el 90 por ciento de las colonias: todas aquellas cuyos plásmidos no transportaban ADNc y que, por tanto, no podían codificar interferón.

En el segundo paso debíamos identificar, en el 10 por ciento de colonias restante, los plásmidos recombinantes que transportaban ADNc para interferón. Para ello, aplicamos el ARN mensajero procedente de leucocitos inducidos a varios lotes de ADN del plásmido de las colonias. Cualquier ARN capaz de hibridarse con el ADN de un plásmido se separaba del ADN; se valoraba, en ovocitos, su actividad de interferón. Posteriormente, en los lotes positivos, se examinaban todos los clones, uno a uno, de la misma manera.

Por este método encontramos un recombinante, el plásmido 104, que contenía la mayor parte de la secuencia del ADN que determina el interferón alfa humano. Encontramos otro, el plásmido 101, que transportaba la secuencia para el interferón beta humano. (Esta sí fue una sorpresa: su ARN mensajero se había sintetizado en leucocitos y no en fibroblastos. Ahora sabemos ya que un determinado tipo celular puede fabricar más de una clase de interferón.) Por esas fechas, otros grupos de investigadores anunciaban la clonación de ADN de interferón humano. Shigekazu Nagata y Charles Weissmann, de la Universidad de Zurich, y sus colegas acababan de clonar un ADN de interferón alfa que no coincidía con el nuestro. El gen del interferón beta fue clonado por Tadatsugu Taniguchi, de la Fundación Japonesa para Investigaciones sobre el Cáncer (quien había anunciado su éxito unos meses antes que el resto de nosotros), por Walter Fiers, de la Universidad estatal de Gante, por Norman H. Carey y otros, en la G. D. Searle and Co. Ltd., y también por nosotros en colaboración con investigaciones de la firma Genentech, Inc.

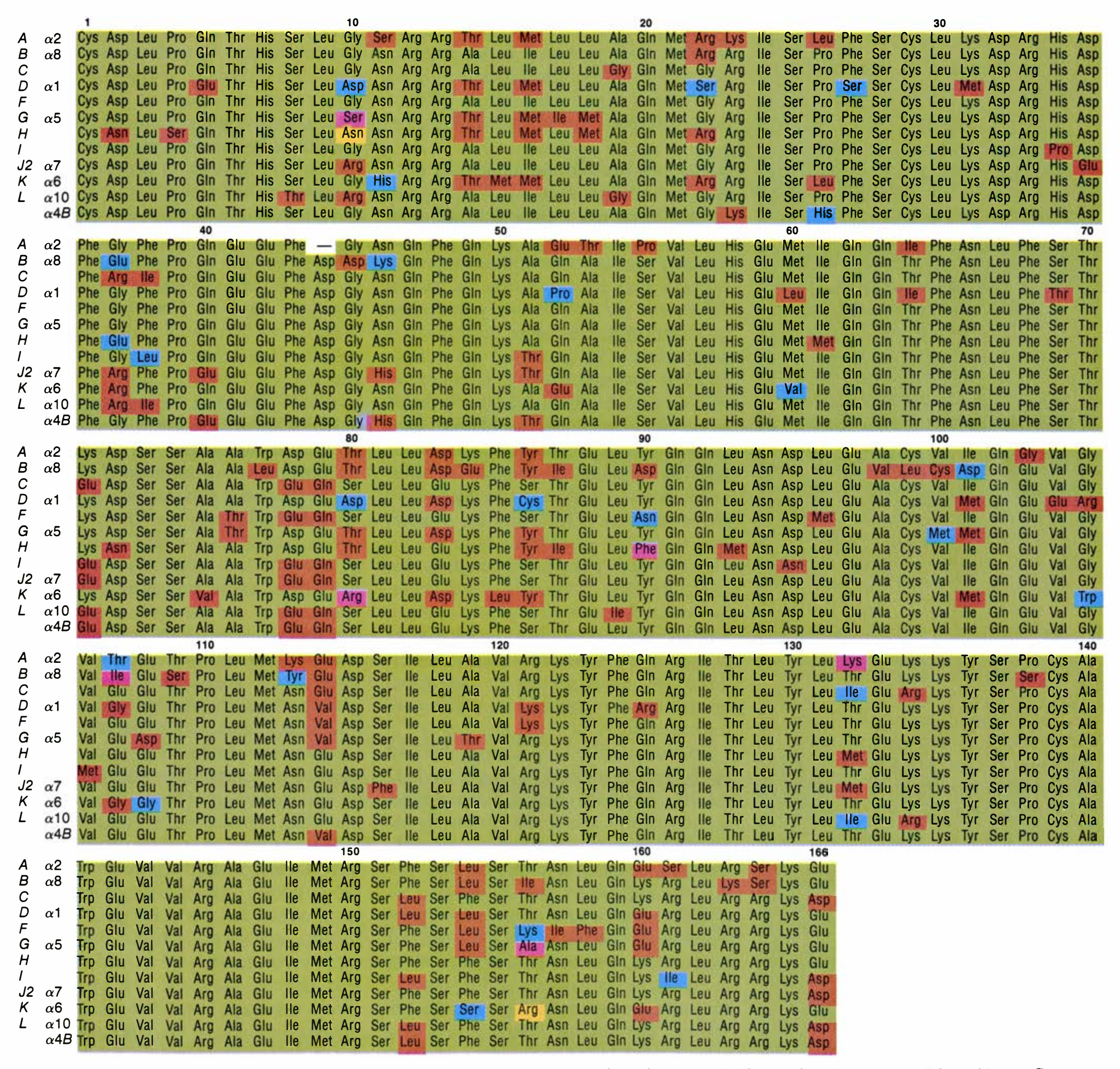
El plásmido 104 nos sirvió de sonda para buscar en una biblioteca estandarizada de ADN humano y encontrar genes de interferón emparentados. Llevamos aisladas 16 secuencias distintas de interferones alfa humanos, a las que hay que agregar las aisladas por otros investigadores. Ninguno de los ADN de interferón contiene secuencias intercaladas (intrones), segmentos no codificantes de ADN que interrumpen las regiones que codifican las proteínas en la mayoría de los genes de los mamíferos. Las secuencias de nucleótidos de los ADN confirman lo que ya habían indicado las secuencias de aminoácidos de los interferones purificados: que los interferones alfa son una clase de proteínas emparentadas, cada una de las cuales está codificada por una familia de genes relacionados.

El plásmido 104 no incluye la secuencia completa que determina un interferón alfa; debido a los caprichos inherentes a la construcción de recombinantes, falta el principio del gen. En Genentech, David V. Goeddel y sus colaboradores utilizaron el plásmido 104 como sonda para buscar un gen entero en la biblioteca de ADNc. Para lograr la expresión de este nuevo gen en *E. coli* había primero que reconstruirlo. Goeddel eliminó en él un segmento que codificaba el líder, un péptido que indica a la célula que debe segregar la proteína, y que se elimina durante la secreción. Roberto Crea añadió un segmento de ADN que transportaba un codón de iniciación, una señal formada por tres nucleótidos necesaria para traducir el ARN mensajero a proteína. El segmento se insertó en el gen, delante mismo del comienzo del segmento que codifica los aminoácidos del interferón. Por último, los investigadores de Ge-

nentech agregaron, antes del codón de iniciación, una secuencia reguladora que contenía la señal propia de la bacteria para empezar la transcripción del ADN a ARN mensajero. Se obtuvo así un gen quimérico que combinaba una región reguladora bacteriana con la región codificadora del interferón del gen humano.

Al insertar en *E. coli* el gen quimérico reconstruido, la bacteria empezó a sintetizar grandes cantidades de interferón humano; tanto, por litro de cultivo bacteriano, cuanto hubieran producido los leucocitos de 100 donantes de sangre. El interferón *A* recombinante de leucocito humano, como lo denominamos nosotros, se comportaba igual que su equivalente fabricado por células humanas. En uno de los primeros experimentos con animales, realizado por Noel Stebbing, en Genentech, este interferón impidió la adquisición de una encefalomiocarditis, una enfermedad vírica, en monos ardilla. Sin embargo, antes de que el producto bacteriano pudiese recetarse a humanos, fue menester realizar una rigurosa purificación para eliminar las proteínas bacterianas contaminantes. Recurrimos a los anticuerpos monoclonales.

Un anticuerpo es una proteína del sistema inmunitario que reconoce y se une a una proteína extraña, o antígeno. Desde 1975 se vienen preparando grandes cantidades de anticuerpos monoclonales: anticuerpos dirigidos contra antígenos específicos. Puesto que disponíamos de nuestro interferón purificado, Theophil Staehelin, de F. Hoffmann-La Roche and Co. Ltd., preparó anticuerpos monoclonales dirigidos contra moléculas específicas de interferón. Los anticuerpos monoclonales específicos para el interferón alfa se unían a perlas de vidrio empaquetadas en columnas de cromatografía. Para



7. INTERFERONES ALFA cuyos genes se han podido aislar. Forman una familia de proteínas estrechamente emparentadas. Se muestran aquí sus secuencias de aminoácidos, deducidas a partir de la secuencia de nucleótidos del ADN que las codifica; se han dispuesto en hilera y se han coloreado para mostrar un alto grado de homología. Las especies cuyos genes fueron aislados por investigadores del laboratorio del autor, en colaboración con Genentech Inc., se indican con letras mayúsculas, a la izquierda. Los genes idénticos, o estrechamente relacionados, secuenciados por el grupo de Charles Weissmann se señalan con una *alfa* en la columna siguiente. La secuencia que se muestra aquí es la que se ajusta a la versión del autor, salvo en α5 y α*4B*.

purificar el interferón bacteriano, Staehelin, Hsiang-fu Kung y Donna S. Hobbs vertieron sobre la columna un extracto de células de *E. coli* que habían sintetizado interferón *A* recombinante. El interferón, y solamente el interferón, se unía a los anticuerpos; los otros componentes, toxinas bacterianas incluidas, atravesaban la columna. Posteriormente se hacía pasar una solución ácida para eluir el interferón, virtualmente puro.

La disponibilidad de grandes cantidades de interferón muy puro abrió la puerta a los ensayos clínicos, que describiré a continuación. Ello nos permitía también cristalizar el interferón, el primer paso para el análisis de la estructura tridimensional de la proteína por cristalografía con rayos X. David L. Miller y Kung prepararon cristales de interferón *A* recombinante humano.

Los ensayos clínicos anteriores estaban limitados, como he mencionado, por la escasez y el elevado costo del interferón sintetizado por células humanas y por las impurezas de las preparaciones que empañaban la valoración del efecto del interferón, y limitaban las dosis utilizables (pues no podíamos administrar a los pacientes gran cantidad del material crudo). La purificación en gran cantidad del interferón alfa recombinante humano eliminó esos impedimentos. Después de realizar las pruebas de seguridad pertinentes, utilizando animales, el producto bacteriano preparado por Hoffmann-La Roche, Inc. recibió la aprobación para su empleo en ensayos con seres humanos. En enero de 1981, Jordan U. Gutterman, del M. D. Anderson Hospital, inició un ensayo clínico encaminado a establecer la seguridad, toxicidad y efectos secundarios de varios niveles en sangre del interferón *A* recombinante en pacientes con cáncer.

Son más de 500 los pacientes que han recibido interferón en pruebas realizadas por varias instituciones académicas y coordinadas por Zofia Dziewanowska, de Hoffmann-La Roche, y sus colegas. Los efectos secundarios más frecuentes coincidían con los advertidos anteriormente en los ensayos con interferón crudo: fiebre, escalofríos, dolores musculares, trastornos gastrointestinales leves, fatiga y anorexia. Los pacientes parecían desarrollar tolerancia a los efectos secundarios "parecidos a una gripe", pero la fatiga y la anorexia aumentaban con la dosis y la duración del tratamiento. Había también un descenso del número de glóbulos blancos de la sangre, que revertía a los pocos días, y en unos cuantos casos se observó una elevación de los enzimas del hígado, que no parecía tener efectos nocivos.

Algunos de los estudios no sólo iban encaminados a valorar los efectos secundarios, sino que intentaban evaluar la respuesta antitumoral. La mayoría de los pacientes presentaban cánceres avanzados, resistentes a la terapia convencional. Aún dentro de ese grupo, se observaron algunas regresiones de los tumores en cáncer de riñón, melanoma múltiple maligno y algunas otras dolencias malignas. A tenor de ciertos resultados, el interferón *A* recombinante sería eficaz contra el sarcoma de Kaposi y en pacientes con el síndrome de deficiencia inmunitaria adquirida (AIDS). El cáncer de mama, de pulmón y de colon mostraban respuestas mínimas o nulas. En todo caso, se trata de resultados provisionales, que no permiten extraer conclusiones definitivas. Se están realizando ya más ensayos de la eficacia del interferón *A* en el tratamiento del cáncer y de enfermedades víricas.

También puede mejorar, gracias al interferón, la detección de tumores. Se sabía que altera la composición proteica de la superficie de la célula. Recientemente, Paul B. Fisher, I. Bernard Weinstein y Soldano Ferrone, de la Facultad de Medicina y Cirugía de la Universidad de Columbia, Jeffrey Schlom y John W. Greiner, del Instituto Nacional del Cáncer, y el autor han observado que el interferón *A* recombinante incrementa la expresión de ciertas proteínas asociadas a los tumores en las células del melanoma maligno y del cáncer de mama. Este efecto quizá permita, con ayuda de anticuerpos monoclonales para estos antígenos tumorales, diagnosticar un cáncer en desarrollo con bastantes meses de antelación.

El *A* no es más que el primero de los interferones bacterianos sometidos a prueba. Se va disponiendo ya de otros interferones alfa para someter a ensayo y a ellos seguirán los beta y gamma. Algunas clases, especies y combinaciones de interferón quizá sean más eficaces que otras frente a determinadas enfermedades y bajo determinadas condiciones. Es más, el médico no se verá limitado a un grupo de interferones naturales. Pueden romperse los genes de interferón e insertar los trozos para fabricar nuevos genes que se traduzcan en interferones híbridos. Al igual que otros investigadores, hemos experimentado ya con esas moléculas híbridas. Con un conocimiento mejor de los mecanismos de las actividades de esas proteínas resultará viable fabricar moléculas de interferón a medida, que ejerzan en grado óptimo determinados efectos.

Campos magnéticos en el cosmos

El mecanismo de la dinamo para la generación de campos magnéticos explica por qué Venus no tiene campo, por qué el Sol lo posee oscilante y por qué el campo galáctico dominante es paralelo al plano del disco

E. N. Parker

Si las fuerzas nucleares y gravitatorias fuesen las únicas que actuasen en el universo, el cuadro completo de la evolución cósmica sería el de una degradación térmica gradual salpicada de sucesos explosivos ocasionales. El cosmos se asemejaría a los serenos –y monótonos– cielos de la concepción clásica. Existe, sin embargo, un agitador cósmico: el campo magnético. Aunque sólo una pequeña parte de la energía disponible en el universo se invierte en los campos magnéticos, éstos son responsables de la mayor parte de la continua actividad violenta que se registra en el cosmos, desde las auroras polares en la atmósfera terrestre hasta las fulguraciones estelares y emisiones de rayos X, pasando por el amontonamiento de nubes de gas interestelar en las galaxias.

Dentro del sistema solar, los vehículos espaciales han transportado magnetómetros hasta las proximidades de todos los planetas, de Mercurio a Saturno. Mercurio, la Tierra, Júpiter y Saturno tienen campos magnéticos propios, no así Venus ni Marte. La intensidad de los campos varía desde $3{,}5 \times 10^{-3}$ gauss en los polos (Mercurio) hasta 8 gauss en los polos (Júpiter). En comparación, el campo magnético terrestre vale 0,6 gauss en los polos. El vehículo espacial *Voyager II*, que transporta un magnetómetro, debe pasar por la proximidad de Urano en 1986. En virtud de su semejanza con Júpiter y Saturno, se espera que haya campos magnéticos tanto en Urano como en Neptuno.

Salvo contadas excepciones, los campos magnéticos de las estrellas no se pueden detectar directamente desde la Tierra, pero sí deducir su existencia a partir de la presencia de actividad semejante a la actividad solar magnética. En razón de este dato, parece que la mayoría de las estrellas poseen campos magnéticos tan intensos, al menos, como el del Sol. Hay algunas clases de estrellas cuyos campos magnéticos alcanzan una intensidad tal que permite su medición directa, al discriminar, por ejemplo, hasta qué punto los dos estados de polarización de una línea espectroscópica dada se apartan de la posición de la línea no polarizada. Las estrellas magnéticas *A* tienen campos de hasta 34.000 gauss; los de algunas estrellas enanas blancas son de 10^7 a 10^8 gauss; los pulsares, de 10^{12} gauss. El disco de nuestra galaxia está atravesado por un campo de intensidad entre 2×10^{-6} y 3×10^{-6} gauss; otras galaxias parecen tener campos de pareja intensidad cuando menos. Hay indicaciones, aunque se trata de algo controvertido, de que un débil campo magnético llena el espacio entre las galaxias.

Y lo que resulta sorprendente: existe una sola explicación genérica de la capacidad de cuerpos tan diferentes como un planeta frío y denso y un disco galáctico tenue y caliente para generar un campo magnético. La explicación, elaborada primeramente para la Tierra, se toma prestada de la magnetohidrodinámica, o estudio de la interacción de un fluido móvil y buen conductor y un campo magnético.

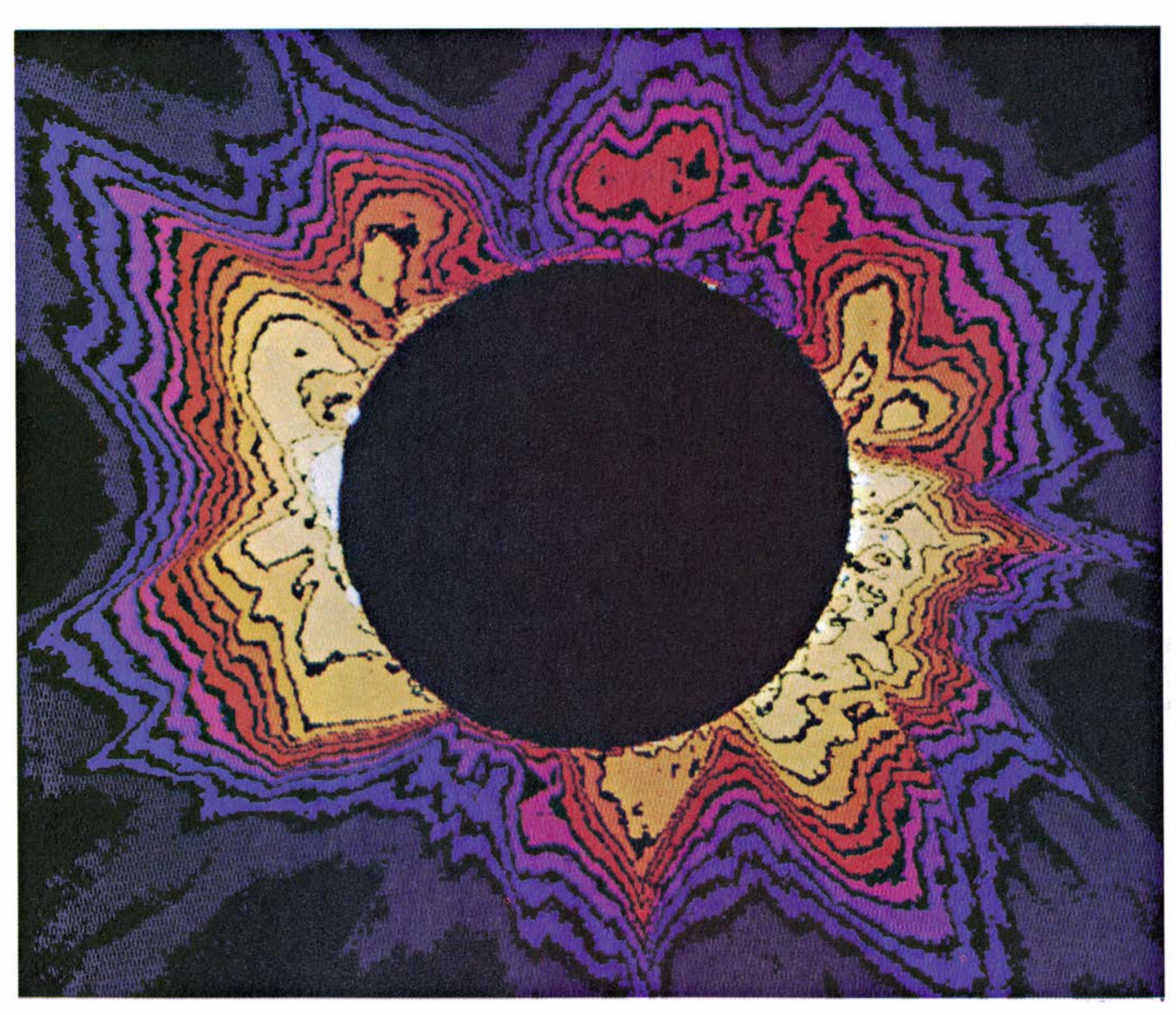
El cosmos está lleno de fluidos capaces de transportar corrientes eléctricas. La mayoría de ellos son gases calientes e ionizados, pero en los planetas hay depósitos internos de metales fundidos. La energía liberada en el interior de los planetas y estrellas y en conjuntos de estrellas por la acción de las fuerzas nucleares y gravitatorias mantiene esos fluidos en movimiento turbulento. Los campos magnéticos que acompañan a los fluidos se estiran y retuercen por el movimiento de estos últimos, ganando energía en el proceso. Dicho de otra manera, los fluidos turbulentos funcionan cual dinamos, aparatos que convierten energía mecánica en energía de los campos magnéticos. La capacidad de los campos magnéticos para alimentarse (como lo hacen los organismos vivientes) de la corriente general de energía procedente del interior de los planetas, estrellas y galaxias, explica por qué florecen dentro y alrededor de casi todos los cuerpos celestes.

Los modelos matemáticos de dinamos dotados de varias características físicas simulan de manera convincente los campos producidos por cuerpos astronómicos. El modelo de capa de la zona convectiva del Sol, por ejemplo, predice un campo creciente y menguante que invierte su polaridad a intervalos regulares, como hace el campo del Sol. Ante la falta de información sobre determinados factores, los movimientos del gas en el interior del Sol entre ellos, los modelos de dinamo son inevitablemente imperfectos. En algunos casos, la información disponible plantea cuestiones para las que no hay respuesta todavía. Un ejemplo es la aptitud de Mercurio para generar un campo. Además, el mecanismo de dinamo, por sí mismo, no puede dar cuenta de los campos excepcionalmente intensos de algunas estrellas. A causa de estos vacíos de información, no se puede refutar la hipótesis rival de que hay campos primitivos, aprisionados, por ejemplo, en el núcleo estable de las estrellas desde su formación. No obstante, una ponderación juiciosa de las pruebas existentes indica que los planetas, el Sol, la mayoría de las estrellas y la galaxia funcionan como colosales dinamos.

Una dinamo convierte la energía del movimiento de un conductor en energía de una corriente eléctrica y de un campo magnético. Una sencilla dinamo de laboratorio consta de un disco de metal que gira en torno a su eje y encima de una bobina conductora cuyo eje está alineado con el del disco. La bobina está eléctricamente conectada al disco y al eje de éste mediante escobillas. Una corriente que pasa por la bobina induce un campo magnético alineado con el eje de la bobina. En el

disco, los electrones, al moverse en el seno de este campo, quedan sometidos a una fuerza perpendicular a su dirección de movimiento y a la dirección del campo magnético, esto es, una fuerza dirigida a lo largo del radio del disco. La dirección de la fuerza viene dada por la conocida regla de la mano derecha. Si el disco, visto desde arriba, está girando en sentido antihorario y el campo magnético está dirigido hacia arriba, la fuerza electromotriz inducida produce una corriente que fluye desde el eje hacia la periferia del disco. La corriente circula entonces desde el disco, a través de las escobillas, hacia la bobina, amplificando el campo magnético generado por ésta, lo que a su vez intensifica la corriente que circula por el disco. Las dinamos magnetohidrodinámicas funcionan de acuerdo con el mismo principio, aunque los electrones no se hallan confinados a cables arrollados y pueden moverse en el seno de un volumen de fluido. La explicación de cómo un fluido conductor móvil intensifica un campo magnético se complica si se prosigue el razonamiento en términos de fuerzas electromotrices inducidas, corrientes resultantes y campos magnéticos asociados con las corrientes.

Son útiles aquí algunos atajos conceptuales. El campo magnético se puede imaginar constituido por líneas de fuerza, es decir, por curvas cerradas a lo largo de las cuales se alinearía una aguja magnética. La intensidad del campo en cualquier región dada del espacio se puede representar esquemáticamente por el número de líneas que atraviesan dicha región, o, de manera menos corriente, por el espesor de las líneas. Siguiendo una idea propuesta por Hannes Alfvén, uno de los fundadores de la magnetohidrodinámica, podemos suponer que las líneas de intensidad del campo están "congeladas" en el interior del fluido conductor o "ligadas" a las partículas que lo integran. El campo se mueve con el fluido y las líneas de campo son deformadas por los movimientos del fluido, como si fueran cintas de goma. Si las partículas a las que está ligada una línea de campo se mueven perpendicularmente a dicha línea y a diferentes velocidades, la línea



1. MAGNETISMO DEL SOL, revelado en esta imagen generada por ordenador de la corona solar. Como el gas de la corona está ionizado, la distribución de sus átomos eléctricamente cargados refleja la forma del campo magnético solar. Esta imagen en falso color es un análisis realizado por ordenador de una fotografía de la corona hecha en la India durante el eclipse solar total del 16 de febrero de 1980. La fotografía se debe a un grupo del Observatorio de la Universidad de Colorado y de la Southwestern con una cámara especial desarrollada por Gordon A. Newkirk, Jr. En la cámara se interpone un filtro graduado radialmente entre la lente y la placa fotográfica. El filtro atenúa más luz hacia el centro que hacia la periferia, impidiendo que la corona interior, más brillante, enmascare los detalles del resto. Los colores corresponden a la intensidad de la luz a la longitud de onda de 6400 angström en cada parte de la corona; las regiones blancas tienen la máxima intensidad, y las regiones violeta oscuro la mínima. Las líneas oscuras son curvas de igual intensidad.

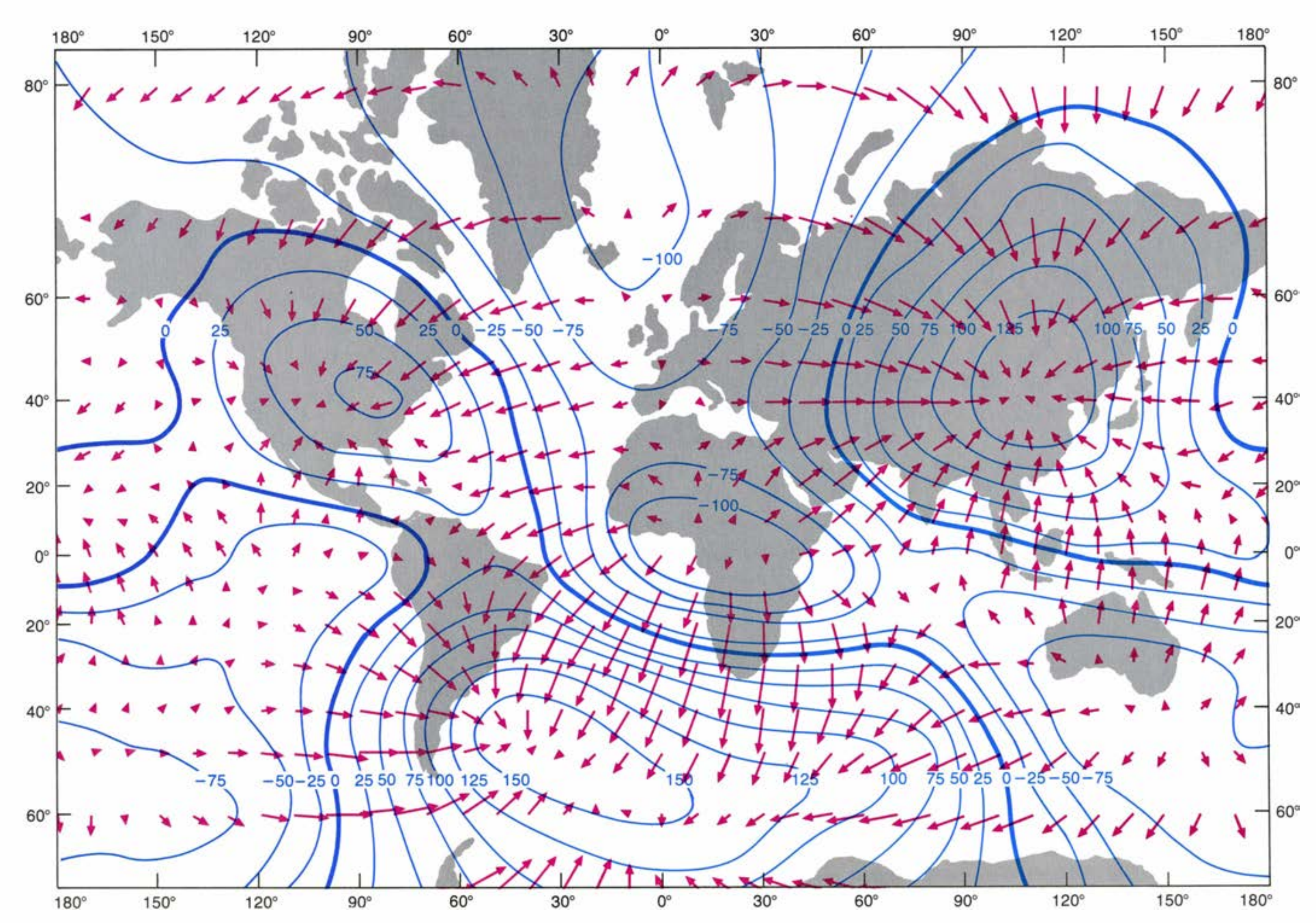
de campo se alargará. De acuerdo con esta interpretación, el alargamiento corresponde a un aumento de la intensidad del campo. La energía del movimiento de las partículas se convierte en energía de un campo magnético, y las fuerzas electromotrices inducidas impulsan la corriente asociada al campo, del mismo modo que lo hacen en la dinamo de disco.

Obsérvese que el funcionamiento de cualquiera de los dos tipos de dinamo requiere la presencia inicial de un débil campo magnético o una débil corriente (la corriente en la bobina en el caso de la dinamo de disco), al menos. Por consiguiente, el mecanismo de la dinamo no explica el origen de los campos magnéticos de los planetas y estrellas, sino cómo se amplifican y mantienen a pesar del constante debilitamiento del campo por la disipación de la corriente asociada. Otros mecanismos que se tratarán después revelan de qué manera pudieron crearse los débiles campos magnéticos que el mecanismo de dinamo intensifica.

La ubicuidad de los campos magnéticos sugiere que las circunstancias bajo las cuales se desarrolla una dinamo magnetohidrodinámica no son especialmente severas. El primer requisito es la presencia de un fluido conductor eléctrico capaz de mantener las corrientes asociadas al campo. El universo está lleno de tales fluidos. La mayoría de los planetas tienen un núcleo metálico fundido; las estrellas están constituídas por gases ionizados, y casi todo el espacio está inundado por un gas con electrones libres suficientes para comportarse como buen conductor de la electricidad. En la mayoría de los casos, el gas no es tan buen conductor como un metal sólido; ahora bien, la capacidad de transporte de corriente de un cuerpo es proporcional a su sección además de serlo a su conductividad; por tanto, la gran extensión del gas lo capacita para transportar grandes corrientes. La atmósfera fría, densa y no conductora y la superficie rocosa y mala conductora de planetas tales como la Tierra son los únicos aisladores eléctricos efectivos que se hallan en el universo.

El segundo requisito se refiere a la configuración del movimiento fluido que habrá de intensificar el campo magnético. En principio, esto lo harían muchos tipos de movimiento fluido, pero la combinación que se produce naturalmente de rotación no uniforme y convección ciclónica es una receta muy eficaz. Lo único que hace falta para tal combinación es un cuerpo en rotación que contenga, o esté compuesto de, un fluido en convección. En virtud de la convección, la rotación tiende a ser no uniforme; como consecuencia de la rotación, la convección es ciclónica.

Un ejemplo familiar de esta clase de movimiento nos lo ofrece la circulación de la atmósfera terrestre. La variación de la dirección de los vientos dominantes con la latitud refleja la rotación no uniforme de la atmósfera. Los alisios tropicales son vientos del Este porque el aire con un determinado momento cinético a altas latitudes pierde velocidad de rotación al moverse a mayor distancia del eje de rotación terrestre; de aquí que en la zona ecuatorial, el aire se mueva de Oeste a Este más lentamente que la superficie terrestre. Las células convectivas de la atmósfera (las zonas de alta y baja presión) adquieren



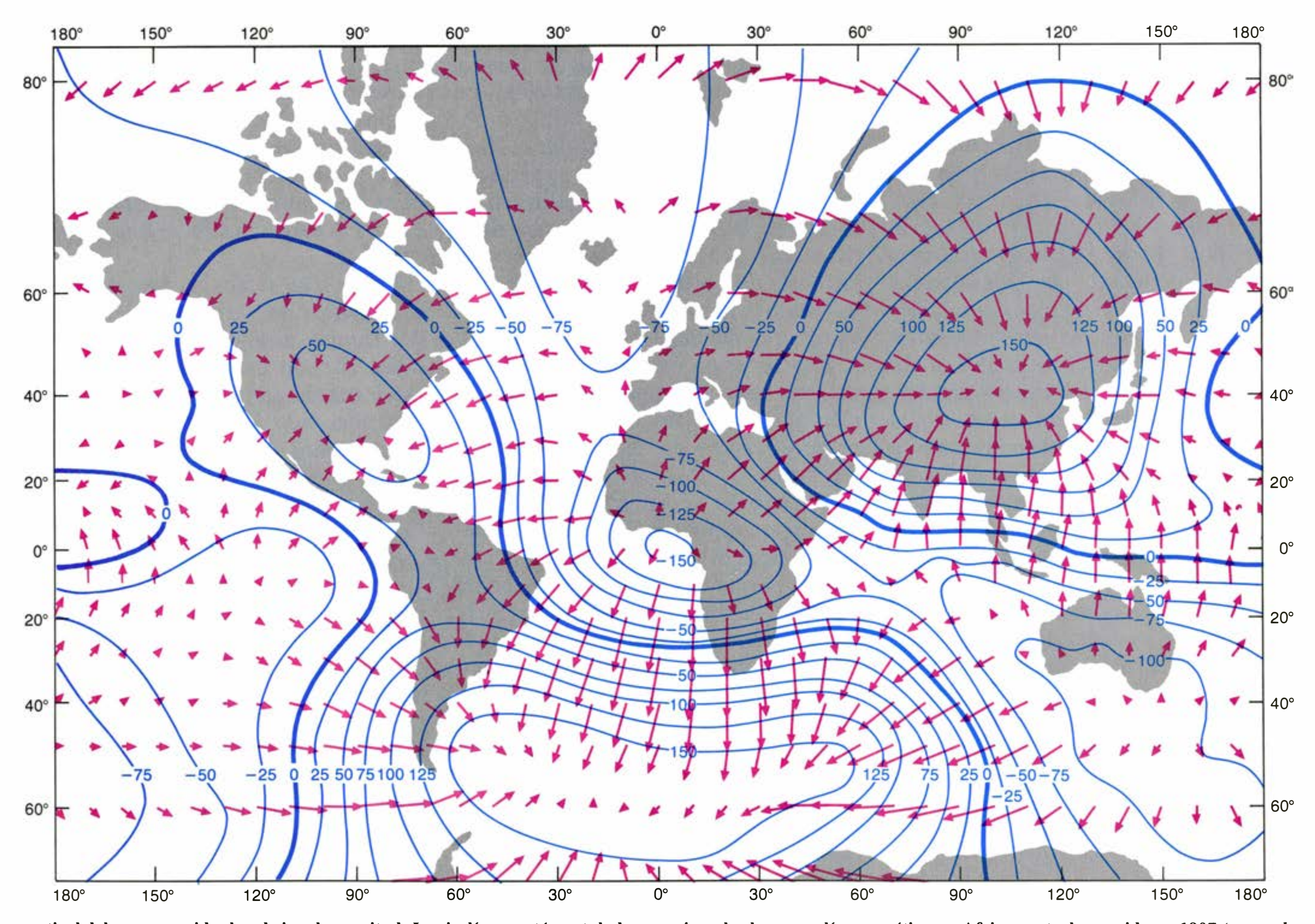
2. ANOMALIAS LOCALES en el campo magnético terrestre. Corresponden, probablemente, a células de convección en el núcleo de la Tierra. Estos mapas muestran la componente residual o irregular del campo terrestre, es decir, la parte que queda cuando del campo observado se resta un campo dipolar en dirección Norte-Sur. Las flechas representan la componente horizontal del campo residual. Las flechas más largas simbolizan una intensidad de campo de unas 10.000 gamma. (Una gamma equivale a 10^{-5} gauss.) Las curvas unen puntos sobre la superficie de la Tierra donde la componente

un giro ciclónico en virtud de la rotación terrestre. Resulta de ello que, en las zonas de ascendencia (baja presión), el aire se mueve girando en sentido antihorario en el Hemisferio Norte y en sentido horario en el Hemisferio Sur; lo contrario sucede en las zonas de descendencia o alta presión. La atmósfera terrestre no se comporta como una dinamo, por la sencilla razón de que no hay en ella electrones libres para transportar la necesaria corriente eléctrica. Sin embargo, pueden esperarse movimientos semejantes en el núcleo en fusión de la Tierra, en la zona convectiva del Sol y, en escala gigantesca, en el disco gaseoso de la galaxia, todos los cuales son suficientemente conductores para comportarse como dinamos.

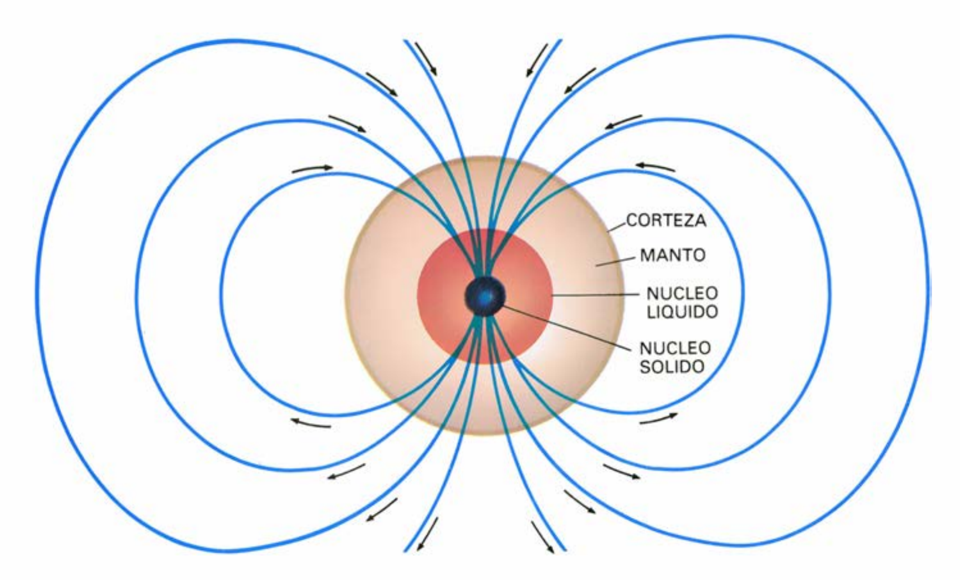
La primera dinamo cósmica de la que se elaboró un modelo satisfactorio fue la terrestre. Una condición previa necesaria para la elaboración del modelo era un conocimiento detallado del interior de la Tierra. Este lo proporcionó el estudio de las trayectorias y velocidades de las ondas sísmicas que parten del foco de un terremoto. Hacia 1940 se sabía que la mitad interior del radio terrestre (el radio es de 6400 kilómetros) está ocupada por metal fundido con una conductividad no mucho menor que la del hierro sólido ordinario. En el centro del núcleo líquido hay un pequeño núcleo sólido de metal cristalino con un radio aproximado de una octava parte del terrestre. El pequeño núcleo sólido no interviene de una forma decisiva en la generación del campo magnético, lo que explica que se desprecie al abordarse el tema. La mitad exterior del radio de la Tierra está constituida por el manto y la corteza. Los minerales silíceos calientes de esas capas son relativamente malos conductores de la electricidad, y a efectos de esta explicación se pueden tratar como aislantes. En este contexto, el manto es en gran parte un impedimento que oculta los movimientos fluidos y las corrientes eléctricas del núcleo y enmascara las fluctuaciones rápidas de pequeña escala del propio campo magnético.

El campo magnético dipolar (de dos polos) detectado en la superficie terrestre está necesariamente asociado a corrientes eléctricas circulares que van de Este a Oeste en el núcleo metálico fundido. Un sencillo cálculo indica que hay corrientes de unos 2×10^9 ampère asociadas con el campo. Al principio, los intentos de explicación del campo magnético terrestre apelaron a múltiples efectos atómicos, el propio efecto termoeléctrico entre ellos, como agentes potencialmente capaces de producir la separación de cargas requerida para crear la necesaria circulación de corriente. En 1945, Walter M. Elsasser indicó que los progresos en física atómica y física del estado sólido permitían descartar el efecto termoeléctrico y todos los demás procesos atómicos como la causa del campo magnético. Insistió en que la única posibilidad que quedaba era que las corrientes que mantienen el campo fuesen inducidas por el movimiento del líquido conductor a través de las líneas de fuerza magnética del propio campo. De hecho, proponía el efecto de dinamo.

De esta explicación del campo terrestre se deducía que la parte más externa del núcleo fluido estaba en movimiento, cosa de la que no existen pruebas directas debido a la interposición del



vertical del campo residual es de igual magnitud. Las isolíneas están rotuladas en cientos de gamma; las cifras negativas que aparecen sobre las mismas corresponden a campo dirigido hacia abajo, a partir de la horizontal. Las anomalías magnéticas derivan hacia el Oeste a razón de 0,18 grados por año. Por ejemplo, la anomalía magnética en Africa central ocurrida en 1907 (*mapa de la izquierda*) había derivado hasta su costa occidental hacia 1945 (*mapa de la derecha*). Dicha deriva hacia el Oeste muestra que la superficie del núcleo fluido, desde la cual emerge, gira más lentamente que el manto suprayacente.



3. DINAMO TERRESTRE que opera en la parte más externa y fluida del núcleo de la Tierra. La naturaleza del núcleo se ha deducido a partir de las trayectorias de las ondas sísmicas que parten de los focos de los terremotos. Se sabe que el núcleo exterior es fluido (es decir, presenta una resistencia despreciable a los esfuerzos tangenciales comparada con su resistencia a la compresión) porque transmite las ondas de presión, pero no las de cizalladura. El núcleo externo consta, principalmente, de hierro que se separó de los otros elementos en el interior de la Tierra al poco de la formación de ésta. Aunque el propio hierro puede ser un imán, a temperaturas tan altas como las del núcleo los materiales ferromagnéticos pierden sus propiedades magnéticas. El magnetismo del núcleo se debe a las corrientes eléctricas que circulan a su través, no al ferromagnetismo del hierro. El hierro líquido tiene una conductividad no mucho menor que la del sólido, lo que implica que el núcleo conduce las corrientes eléctricas asociadas al campo magnético.

manto. Ahora bien, la variación del campo magnético terrestre con el tiempo se puede explicar sólo por movimientos en el núcleo. Existen registros de la dirección y magnitud del campo en la superficie terrestre desde los tiempos de Carl Friedrich Gauss, gran matemático y físico del siglo XIX cuyo nombre lleva la unidad de inducción magnética. Estos registros muestran una docena o más de anomalías locales identificables en el campo, anomalías que tienen dimensiones de varios miles de kilómetros y amplitudes del orden del 10 por ciento del campo dipolar. Las anomalías varían lentamente con el tiempo, aumentando o disminuyendo a lo largo de ciclos de vida de varios siglos. Elsasser indicó que la componente variable del campo magnético correspondía a una amplia configuración convectiva de una docena de células identificables en el núcleo líquido.

Las células conocidas están situadas en latitudes medias y bajas. Muestran una firme tendencia a derivar hacia el Oeste a razón de unos 0,18 grados por año. La deriva sugiere que la porción superficial del núcleo, a través de la cual emerge el campo, gira más lentamente que el manto situado encima de ella. La velocidad de deriva corresponde a una velocidad del fluido de unos 0,3 milímetros por segundo (alrededor de un metro por hora) en la superficie del núcleo.

Elsasser observó que la explicación más sencilla de la lenta rotación del núcleo residía en la acción de la fuerza de Coriolis sobre el fluido que se eleva y desciende en las células convectivas. La conservación del momento cinético requiere que la velocidad angular del fluido ascendente decrezca al moverse a mayor distancia del eje de rotación de la Tierra. Se deduce que la superficie del núcleo gira más rápidamente a altas latitudes que a bajas latitudes, y que, en éstas, la parte interna del núcleo gira más rápidamente que la superficie.

Elsasser se propuso calcular el efecto del movimiento supuesto del fluido eléctricamente conductor del núcleo sobre el campo del dipolo magnético. La primera observación importante que surgió de su investigación fue que el campo magnético primario del núcleo terrestre era un campo Este-Oeste, perpendicular a la componente principal del campo en superficie. Se le llamó campo azimutal. En general, la parte del campo que yace en planos perpendiculares al eje de rotación de un cuerpo magnético se llama campo azimutal, y campo meridiano a la parte situada en planos que pasan por el eje de rotación. El campo azimutal del núcleo terrestre se origina por el alargamiento de las líneas de fuerza en sentido Norte-Sur del campo dipolar al ser transportadas por el fluido en rotación del núcleo. La sección de una línea de campo que está cerca del eje es transportada más rápidamente que los segmentos más alejados del mismo. La rotación no uniforme alarga las líneas Norte-Sur del campo en dirección Este-Oeste. El campo resultante está dirigido hacia el Este en el Hemisferio Norte y hacia el Oeste en el Hemisferio Sur.

A medida que el giro arrastra las líneas del campo, el campo azimutal se va agrandando. La intensificación continúa hasta que queda compensada por un factor contrario: la tensión de las líneas de fuerza magnética o la disminución resistiva de la corriente eléctrica asociada. Edward C. Bullard, de la Universidad de Cambridge, dándose cuenta de la importancia del trabajo de Elsasser, emprendió el estudio del problema y mostró que el campo azimutal en el núcleo podía centuplicar varias veces la intensidad del campo dipolar observado en superficie. El campo dipolar vale aproximadamente 0,3 gauss en la superficie en el Ecuador, el campo azimutal en el núcleo puede ser de 100 gauss o más.

El campo dipolar en la superficie de la Tierra debía entonces ser un efecto secundario del campo azimutal. El razonamiento era el siguiente: el afloramiento del fluido en las células convectivas del núcleo deformaba las líneas de fuerza del campo azimutal, dando lugar a bucles de campo en los planos meridianos del núcleo que representaban un campo dipolar neto. El mecanismo parecía prometedor, pero demostrar que el movimiento convectivo generaba un campo dipolar constituyó una empresa titánica.

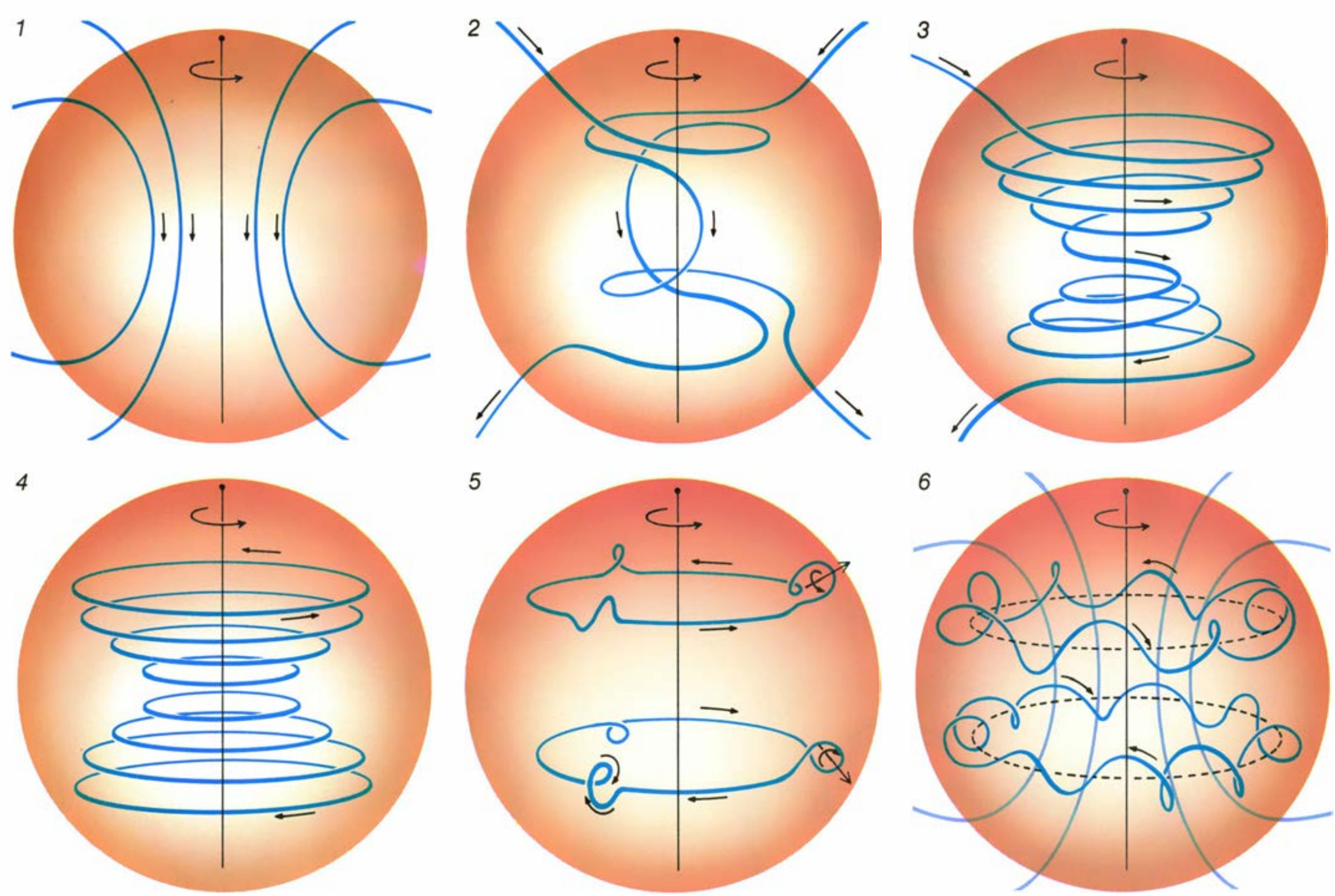
Había que idear una representación matemática del campo de velocidad y del campo magnético que permitiera tratar eficazmente las inacabables interacciones de ambos, y diseñar un modelo adecuado de los movimientos fluidos. Bullard creó una representación de los campos que facilitaba los cálculos. Introdujo luego las velocidades fluidas que podrían esperarse de unas pocas células convectivas estacionarias y aproximadamente circulares. Sus cálculos mostraron que la convección producía distintos campos magnéticos en pequeña escala, con los que a su vez interaccionaba la convección, generando campos de escalas todavía menores. Continuó los cálculos hasta escalas progresivamente menores, sin conseguir pruebas concluyentes de la creación de un campo dipolar neto en gran escala.

En el decenio de 1930, T. G. Cowling, de la Universidad de Leeds, quien había estado estudiando la interacción de un fluido conductor y un campo magnético, había demostra-

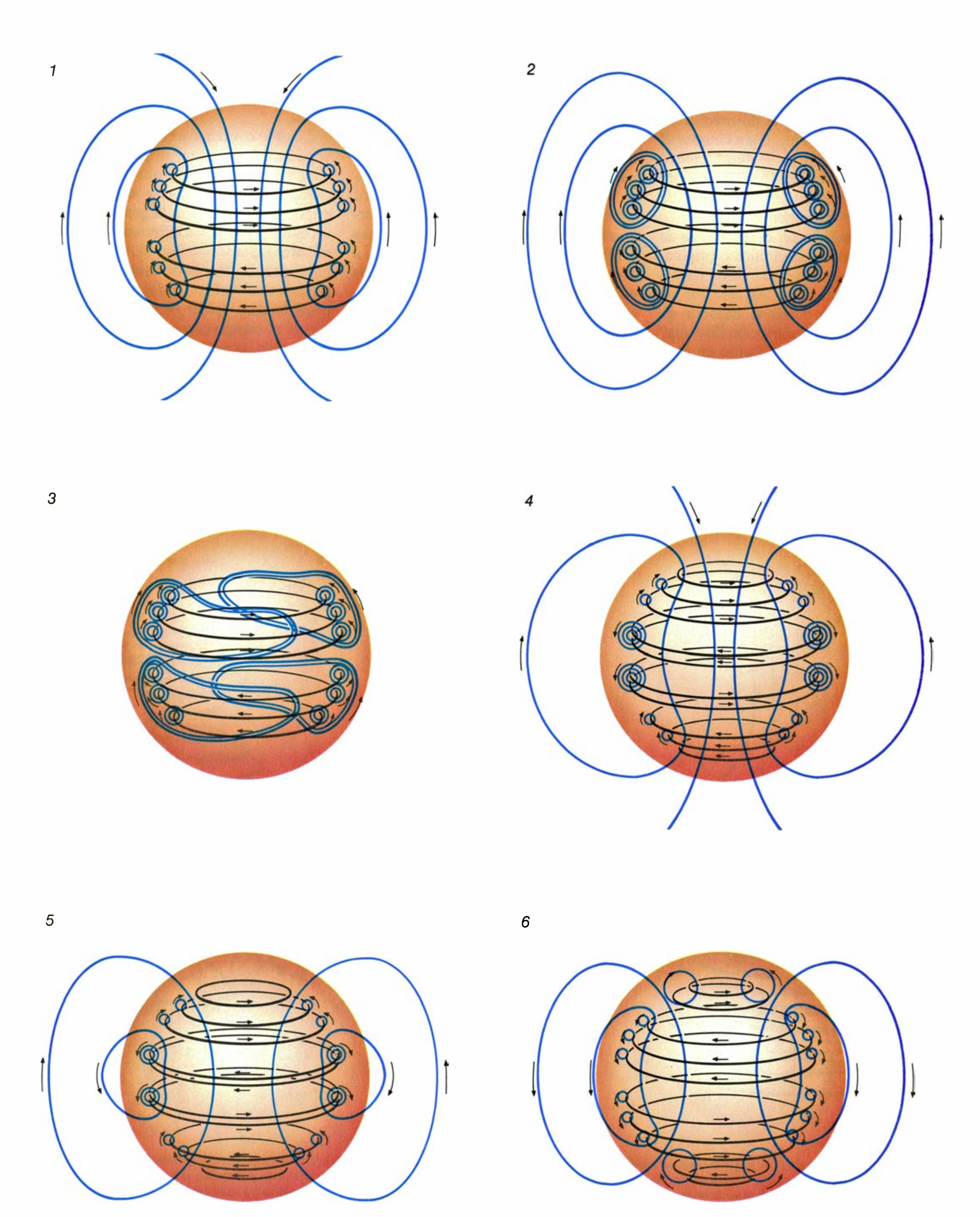
do un teorema que establecía que el movimiento fluido no puede generar un campo dipolar perfecto o cualquier otro campo con simetría de revolución alrededor de un eje. Conociendo el teorema de Cowling, Bullard colocó sus células convectivas irregularmente alrededor del núcleo, en la hipótesis de que el campo toscamente axisimétrico de la Tierra pudiera ser el residuo de un campo asimétrico en el núcleo originado por estos movimientos. Ante los resultados negativos de los cálculos de Bullard, parecía como si hubiera de existir un "super-teorema de Cowling" en el sentido de que un fluido en régimen *estacionario* no podría generar campo magnético alguno, tanto si el movimiento se encontraba acompañado de campos asimétricos como si no. Un "teorema" oscuro era aquel y días de total desorientación eran aquellos para cuantos se interesaban por el origen del campo magnético terrestre.

Tuve la gran suerte de trabajar con Elsasser durante un par de años (al filo de 1954). Bullard y Elsasser mantenían frecuente correspondencia. En cierta ocasión Bullard vino de visita varios días. No pude evitar interesarme por la cuestión del origen del campo magnético terrestre. La conjetura de Bullard sobre la existencia de un "superteorema de Cowling" me produjo una profunda impresión. Comencé, pues, a preguntarme cuál sería el efecto de los movimientos fluidos *no estacionarios* sobre los campos magnéticos. ¿Qué ocurriría si los movimientos ascendentes y descendentes del fluido se pusieran en marcha, breve e intensamente, y después se suprimiesen durante un período largo? Por un lado, los movimientos rápidos y de corta duración simplificarían los cálculos. Cuando el movimiento se interrumpiese, la resistencia eléctrica del fluido haría que el campo "descansara" y se suavizara; los complicados campos en pequeña escala desaparecerían, dejando sólo el residuo en gran escala y el intenso campo azimutal.

Cuando comencé a considerar la convección intermitente, se puso inmediatamente de manifiesto que cuanto más ciclónica fuera la convección, tanto más eficaz se mostraría en la generación de un campo magnético dipolar. El movimiento ciclónico de un fluido eleva y hace girar las líneas de fuerza del campo azimutal, deformándolas hasta constituir hélices. En el Hemisferio Norte, donde el campo corre de Oeste a Este, una célula ascendente crea un bucle de campo cuya parte exterior está dirigida hacia el Norte y cuya parte interior está dirigida hacia el Sur. En el Hemisferio Sur, el campo azimutal tiene la dirección opuesta, pero el giro de las células ciclónicas está también invertido, de modo que la hélice posee la misma dirección de rotación. Cada bucle de una



4. MODELO DE DINAMO de la generación del campo magnético terrestre. Supone que las líneas del campo se alargan por los movimientos del fluido en el núcleo; como resultado, el campo se intensifica a una velocidad suficiente para compensar la velocidad de debilitamiento. Aquí se muestra una versión muy simplificada del proceso. Las líneas de fuerza en sentido Norte-Sur del campo dipolar (*1*) son desviadas hacia la dirección Este-Oeste (*2*) por el fluido en rotación no uniforme del núcleo. La velocidad angular del fluido decrece con la distancia al eje de rotación de la Tierra, y la parte de una línea del campo situada hacia el interior del núcleo es arrastrada por la rotación más rápidamente que la parte que está cerca de la superficie del núcleo. El campo azimutal resultante se dirige hacia el Este en el Hemisferio Norte y hacia el Oeste en el Hemisferio Sur. A lo largo de su vida de 30.000 años, las líneas del campo rodean el núcleo muchas veces y el campo azimutal se va intensificando (*3*). Si el campo dipolar original se resta del campo helicoidal creado por la rotación no uniforme, resulta un campo puramente azimutal (*4*). El afloramiento de fluido en la convección ciclónica crea bucles en los anillos del campo azimutal (*5*). El fluido ascendente describe una hélice en sentido antihorario en el Hemisferio Norte y horario en el Sur. La parte más externa de un bucle en los hemisferios se dirige, en consecuencia, hacia el Norte y la parte más interna hacia el Sur. Cuando los bucles de las hélices en que se ha transformado el campo azimutal se difunden y entremezclan, crean el campo meridiano en gran escala detectado en la superficie terrestre como campo dipolar (*6*).



5. INVERSION DEL CAMPO MAGNETICO TERRESTRE; podría ser producida por un brusco aumento en la velocidad de convección en el núcleo. Ordinariamente, los pequeños bucles meridianos (*azul*), forma en la cual el campo azimutal (*negro*) es transportado por el fluido en afloramiento, tienen tiempo de soldarse, creando el campo dipolar en gran escala (*1*). Si la velocidad de convección aumenta (*2*), los campos meridianos locales no tienen tiempo de unirse antes de que la rotación no uniforme del núcleo comience a alargarlos en dirección Este-Oeste. El nuevo campo azimutal corre en dirección opuesta a la del campo existente en las bajas latitudes (*3*). La suma y anulación de campos traslada el antiguo campo azimutal a latitudes más altas y produce un campo azimutal invertido a latitudes más bajas (*4*). La convección produce bucles meridianos invertidos a partir del campo azimutal inverso. El campo general meridiano creado por la difusión de esos bucles enmascara el campo original (*5, 6*). El esquema que aparece aquí es sólo un modelo.

hélice tiene una pequeña componente meridiana. Cuando los bucles se difunden y se sueldan con otros vecinos durante el tiempo en el que el movimiento del fluido ha cesado, crean una circulación general meridiana del campo que aparece como el campo dipolar. Así, la convección ciclónica intermitente genera un campo dipolar neto.

En 1958, Arvid Herzenberg, de la Universidad de Manchester, demostró que un campo magnético podía originarse por movimientos fluidos estacionarios o por movimientos intermitentes. La conjetura de Bullard resultó, pues, falsa, aunque sirvió para dirigir la investigación en una dirección fructífera. Había encontrado una dificultad práctica, no fundamental: los movimientos fluidos que eligió producían un efecto de dinamo relativamente débil y las correspondientes ecuaciones de dinamo eran, en consecuencia, difíciles de resolver. Desde entonces se ha demostrado que los movimientos fluidos tienen al menos alguna capacidad de generación de un campo magnético; por paradójico que parezca, las únicas excepciones a esa regla son los sencillos movimientos simétricos alrededor de un plano o de un eje, es decir, los movimientos que uno elegiría en busca de las soluciones matemáticas más simples de las ecuaciones de la dinamo. El ingrediente esencial para la generación de un campo es que el movimiento del fluido sea helicoidal: el fluido debe girar alrededor de su dirección de movimiento al mismo tiempo que fluye. Así, la convección ciclónica, incluso en ausencia de rotación no uniforme, basta para producir un campo magnético. La combinación natural de convección ciclónica y rotación no uniforme es, sin embargo, tal vez el esquema más eficaz que podría idearse.

El modelo de la dinamo terrestre se contrastaría si podía demostrarse o no su aptitud para intensificar el campo dipolar a una velocidad suficiente capaz de contrarrestar el debilitamiento del campo. El campo magnético en un cuerpo que transporta corriente se debilita en un tiempo característico (el tiempo de relajación magnética resistiva), que es proporcional a la capacidad de transporte de la corriente del cuerpo. Esta es a su vez proporcional a la conductividad multiplicada por el área de su sección recta. Por su tamaño, el núcleo de la Tierra tiene un tiempo de relajación magnética prolongado. Si la generación del campo cesara hoy, habría de transcurrir unos 30.000 años antes de que la intensidad de la corriente eléctrica y del campo magnético se redujesen a la mitad de sus valores actuales.

La intensidad del campo azimutal viene determinada por el número de veces que las líneas de fuerza se arrollan en torno al eje terrestre a lo largo de su vida de 30.000 años. Por tanto, un campo azimutal de 100 gauss o más en el núcleo en el Ecuador puede estar asociado a un campo dipolar de 0,3 gauss en el Ecuador. La velocidad a la que el campo dipolar se origina del azimutal se puede calcular a partir de una velocidad convectiva de un metro por hora en la superficie del núcleo. El modelo de la dinamo terrestre basado en las ligaduras conocidas (la intensidad del campo dipolar, la resistividad del hierro fundido y la velocidad del fluido en la superficie del núcleo) regenera fácilmente el campo a una velocidad equivalente a reemplazar el campo dipolar una vez cada 30.000 años, es decir, a una velocidad igual a la de debilitamiento.

Otra propiedad del campo terrestre que requiere explicación es su capacidad de invertir la polaridad. En 1955, S. K. Runcorn, de la Universidad de Newcastle-upon-Tyne, indicó que la alineación de los granos de óxido de hierro magnético en las lavas señalaba la dirección del campo magnético local en la época de su cristalización. Prosiguió mostrando que este registro paleomagnético indicaba que el campo terrestre había invertido su polaridad bruscamente (en sólo unos 1000 años) a intervalos aleatorios de 100.000 a 10 millones de años. Desde entonces se ha demostrado que cualquier cambio brusco en la posición o en la intensidad de las células convectivas puede arrojar la dinamo a un estado caótico donde los campos locales corran en dirección opuesta y generen un campo en gran escala de polaridad invertida. Los cambios que ello requiere son modestos (unos 20 grados de latitud) y, por tanto, plausibles.

Pese al éxito del modelo hidromagnético de la dinamo terrestre, sería engañoso sugerir que el problema del campo magnético terrestre se ha resuelto por completo. Sólo se ha demostrado que los movimientos fluidos de la clase supuesta en el núcleo de la Tierra podrían generar, en principio, el campo magnético observado. Sin embargo, los movimientos fluidos usados en la demostración matemática formal, e incluso los empleados en los complicados experimentos numéricos recientemente realizados, se habían simplificado de una u otra manera con el fin de hacer tratables los cálculos. El objetivo final es comenzar desde los principios fundamentales, tales como la velocidad de rotación y la fuente de energía que produce la convección. En vez de postular una configuración determinada de la circulación, se podría deducir un modelo cuantitativo preciso de los movimientos del fluido y de los campos magnéticos locales en el núcleo. Los diversos efectos del flujo convectivo sobre el campo y, viceversa, se podrían calcular entonces.

Con toda probabilidad, transcurrirá algún tiempo antes de que el problema se resuelva de manera completa, por culpa, sobre todo, de la falta de información de primera mano sobre la fuente de energía que provoca la convección. Tanto el calor liberado por la desintegración del isótopo radiactivo potasio 40 como la consolidación de una fase más densa del hierro a partir de la fase líquida más ligera intervienen en la convección. De forma semejante, aunque los cambios bruscos en el modo o velocidad de la convección podrían causar cambios de polaridad, no se sabe por qué habrían de producirse tales cambios bruscos. Algunos estudios recientes de sistemas matemáticos acoplados han repetido episodios al azar, pero no se ha evidenciado cómo aplicar esos resultados a la convección en el núcleo.

Si se trata de planetas distintos de la Tierra, la información en que apoyarse escasea mucho más. Así, el examen de los campos planetarios se limita a cuestiones básicas. ¿Por qué Mercurio, un planeta pequeño, tiene un campo magnético propio, mientras que Venus, aproximadamente del tamaño de la Tierra, carece de él? ¿Por qué Marte no tiene campo?

La velocidad a la que un planeta genera un campo magnético viene definida por lo que se conoce como el número de dinamo. Este es el producto de tres magnitudes (la velocidad de giro de las células convectivas, el gradiente de velocidad producido por la rotación no uniforme y el volumen del núcleo) dividido por la resistividad del cuerpo. Si el número de dinamo está por debajo de un valor umbral, el cuerpo es incapaz de generar un campo magnético con la misma velocidad con que éste se debilita. Si el número de dinamo está por encima del umbral, la velocidad de generación excede a la de debilitamiento y el modo de complejidad espacial del campo creciente depende del valor del número. Por ejemplo, un campo dipolar constante viene generado por un

cuerpo con un número de dinamo más bajo que el de un cuerpo que genera un campo cuadripolar oscilante.

Mercurio es un planeta pequeño, aunque denso. De un radio que mide, aproximadamente, un tercio del terrestre, su densidad revela un núcleo metálico bastante grande; el radio de este último viene a ser la mitad del que mide el núcleo terrestre. Mercurio gira con relativa lentitud (con un período de 59 días). A causa de la baja velocidad de rotación, no se esperaba que generase campo magnético. Constituyó una sorpresa, por tanto, que el vehículo espacial *Mariner 10* registrase un campo magnético dipolar al pasar por la vecindad de Mercurio en 1974 y nuevamente en 1975.

Los débiles campos irregulares en la proximidad de Venus y Marte no parecen ser otra cosa que el campo del Sol transportado y aplastado contra esas esferas de roca, no magnéticas, por el viento de partículas eléctricamente cargadas que procede del Sol. Venus, llamado el planeta hermano de la Tierra en razón de sus semejanzas en tamaño y estructura interna, no parece generar campo porque gira muy lentamente (con un período de 244 días). Marte gira esencialmente a la misma velocidad que la Tierra (con un período de 24 horas 37 minutos). La baja densidad de Marte, sin embargo, sugiere que si tiene núcleo metálico, éste es pequeño. La velocidad de disipación de un campo magnético varía en razón inversa del cuadrado del radio del cuerpo magnético; puede suceder, pues, que cualquier campo generado en el núcleo de Marte se pierda antes de que se regenere.

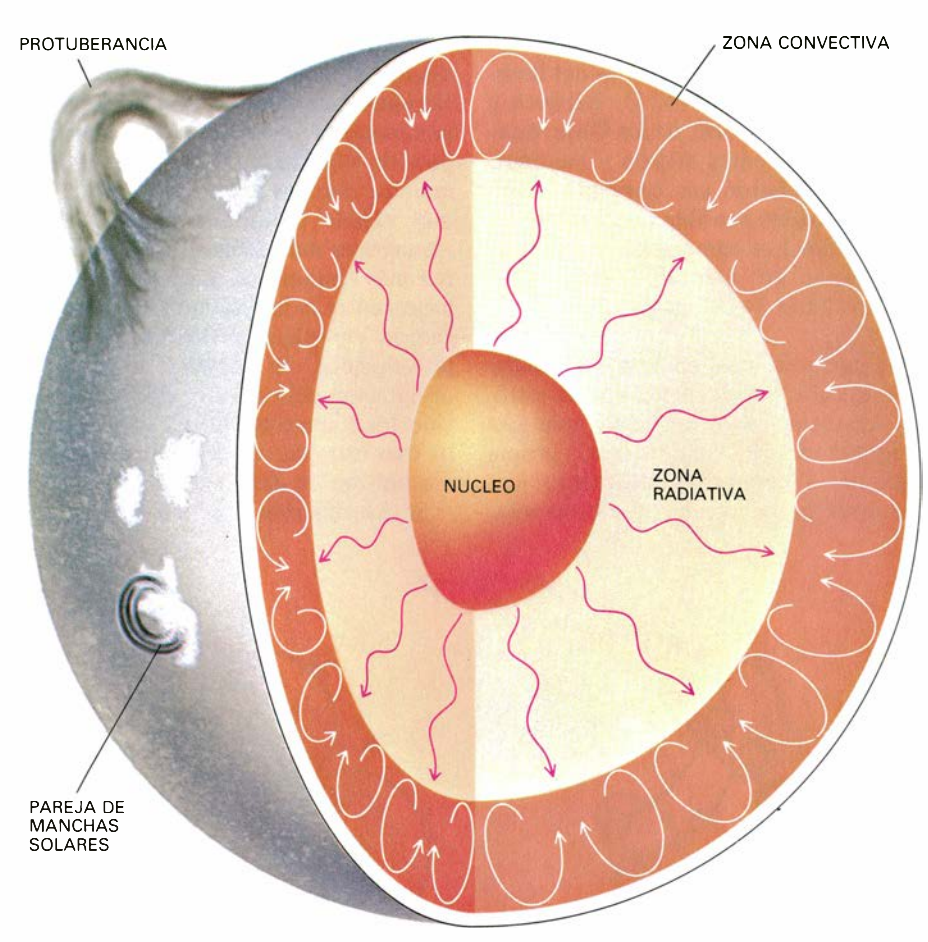
6. INTENSIDAD DEL CAMPO MAGNETICO DE UN PLANETA, correlacionada con el tamaño del núcleo o capa conductora del planeta y con su velocidad de rotación. Venus carece de campo, ya que gira lentamente (con un período de 244 días). Júpiter y Saturno tienen campos intensos porque giran rápidamente (con períodos de unas 10 horas) y poseen cargas interiores de hidrógeno y helio metálicos que son tan conductores como el cobre a la temperatura ambiente. El campo de Saturno en superficie puede ser más débil que el de Júpiter, porque la capa de hidrógeno metálico convectivo de Saturno esté rodeada por una capa estática y quizá rígida de hidrógeno metálico, o porque la convección sea más débil en Saturno que en Júpiter. En vista de la presencia de campo en Mercurio, su ausencia en Marte constituye una suerte de rompecabezas. El tamaño del núcleo de Marte no se conoce bien: puede ser muy pequeño o del tamaño que posee el de Mercurio. Marte gira 50 veces más deprisa que Mercurio y, sin embargo, no genera campo. Puede ocurrir que Marte tenga un núcleo demasiado pequeño para generar un campo magnético o que la convección del núcleo de Mercurio sea intensa y compense la pequeñez de su núcleo.

Las comparaciones entre estos cuatro planetas plantea cuestiones del mayor interés que, por ahora, no tienen respuesta. En la superficie de Mercurio, la intensidad del campo se cifra en una milésima parte de la que posee el campo magnético terrestre en superficie; ocurre, además, que la propia existencia del campo es enigmática. Si Mercurio puede mantener un campo dipolar constante, la Tierra, que gira 59 veces más deprisa y tiene un núcleo dos veces mayor, debería mantener campos más complejos. Quizás algún factor inhiba la generación de tales campos en el núcleo terrestre, o tal vez la convección fluida del núcleo de Mercurio sea más vigorosa que en la Tierra y compense el menor núcleo y la velocidad de rotación, más baja. Marte carece de campo, aunque gira a una velocidad más de 50 veces mayor que Mercurio. Si el núcleo de Marte es comparable en tamaño al de Mercurio, como sostienen algunos investigadores sobre la base de la densidad media del planeta, la ausencia de campo en Marte y la presencia de él en Mercurio es incomprensible. Evidentemente, el núcleo de Marte es menor, o menos convectivo, que el de Mercurio; y quizá sea mucho menor y menos convectivo. Ante la penuria informativa sobre la convección en los núcleos planetarios, preguntas como éstas quedarán probablemente sin contestación por algún tiempo.

La dinamo del Sol es un estrato esférico hueco, no una esfera como en los

núcleos de los planetas. Constituye la zona convectiva del Sol, que empieza justamente debajo de la superficie visible y se extiende una cuarta parte de la distancia al centro del Sol. El período de rotación solar es de 25 a 35 días. Las velocidades de rotación de ciertos rasgos, las manchas solares por ejemplo, muestran que la superficie de la zona convectiva gira de manera no uniforme, teniendo la superficie en los polos una velocidad angular de aproximadamente dos tercios de la velocidad en el ecuador. La convección se deja ver en forma de granulaciones, las marcas en forma de granos de arroz que recogen las fotografías de la superficie solar. Puesto que el Sol está en rotación, se trata de una convección ciclónica.

La dinamo solar funciona de manera muy parecida a la terrestre. El campo azimutal del Sol, a diferencia del terrestre, se observa directamente, pues se ve la cima de la capa convectiva. Las manchas solares emparejadas y las regiones magnéticas bipolares en general son abultamientos locales en el campo azimutal, empujados hasta la superficie del Sol por la fuerza ascensional magnética. El campo azimutal que yace bajo la superficie alcanza una intensidad de 200 gauss por lo menos y llega hasta los 10.000 gauss. El campo meridiano generado a partir del azimutal es visible sólo a altas latitudes, donde aparece en forma de serpentinas coronales y prominencias polares. Es mucho más débil que el campo azimutal, teniendo una intensidad de quizá 5 o 10 gauss.



7. SE LLAMA DINAMO SOLAR a la zona convectiva del Sol, que tiene un radio de aproximadamente una cuarta parte del radio solar. Esta dinamo de estrato delgado no puede generar un campo meridiano con velocidad suficiente para mantener el alargamiento del campo en gran escala en toda la extensión del Sol; el campo es, por tanto, un dipolo oscilante. La manera de actuar la dinamo solar se asemeja a la de la terrestre. Las manchas emparejadas y las fulguraciones solares hablan de intensos campos azimutales locales. Las protuberancias a latitudes altas están alineadas con las líneas del campo meridiano local.

El ciclo magnético de Sol se desarrolla en períodos de 22 años. El reforzamiento y debilitamiento coordinados de los campos azimutal y meridiano en el curso del ciclo hacen que el campo invierta su polaridad aproximadamente una vez cada 11 años. Se puede mostrar que esta danza magnética es una consecuencia de las características de la dinamo solar. Una dinamo de capa delgada como la del Sol pierde campo demasiado rápidamente para que el campo generado alcance equilibrio difusivo a lo largo de toda la extensión (en latitud) de la capa. El campo a baja latitud no ha tenido tiempo de alcanzar el equilibrio antes de que se genere un campo opuesto a altas latitudes. En consecuencia, una dinamo de capa delgada carece de modo dipolar estable; el más bajo modo automantenido que puede existir es un dipolo oscilante.

La explicación de los campos de algunas estrellas magnéticas y de las galaxias a través del mecanismo de dinamo se enfrenta a una hipótesis rival, según la cual los campos son primordiales, aprisionados en las estrellas y en las galaxias desde la formación del universo. Las estrellas se forman por el colapso gravitatorio del gas interestelar. El gas interestelar viene guiado por el campo magnético galáctico, de modo que un cierto volumen de gas lleva consigo parte del campo cuando colapsa para formar una estrella, comprimiendo el campo en cuantía igual a aquella en que se comprime el área de la superficie del gas. Gran parte del campo primordial aprisionado escapa de la estrella en el curso de la vigorosa mezcla que acompaña a las fases finales del colapso, pero es razonable suponer que una proporción apreciable queda retenida.

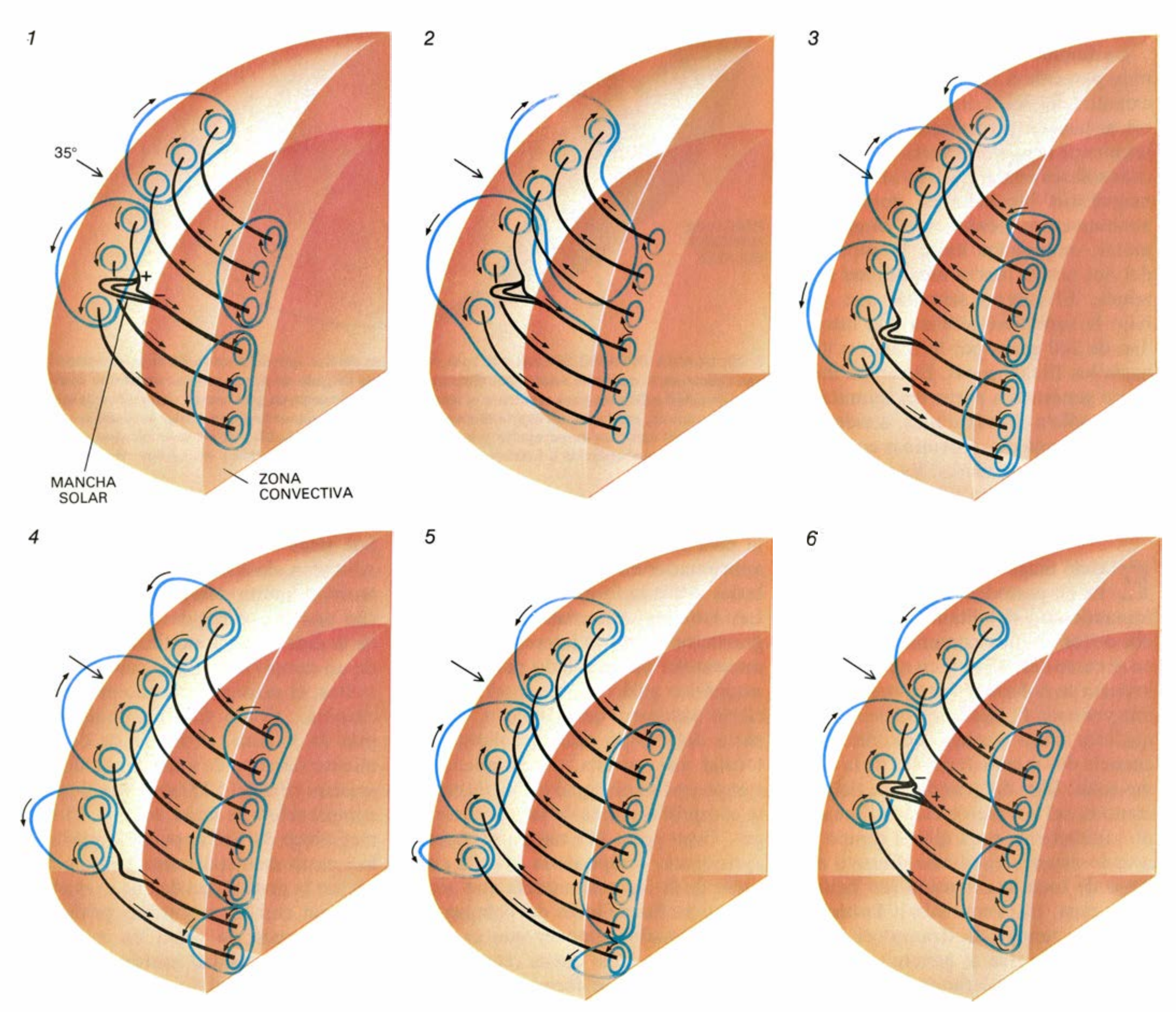
Los planetas se formaron todos de una manera similar; pero en este caso hay un argumento sencillo en contra de los campos primordiales. Como hemos mencionado, el tiempo de relajación magnética del núcleo terrestre es de unos 30.000 años, mucho más corto que la duración de la vida de la Tierra. Esta se formó hacia la misma época que el Sol, unos 4500 millones de años atrás. Abandonado a sí mismo, un campo primordial habría desaparecido hace mucho tiempo; se requiere el mecanismo de la dinamo para explicar la presencia de un campo hoy. En el caso de las estrellas, el razonamiento no es tan concluyente. Como Cowling indicó hace más de 30 años, el tiempo de debilitamiento resistivo de una estrella del tipo solar se cifra en 5000 millones de años, comparable a la edad del Sol. El ciclo magnético del Sol se puede explicar sólo a partir de una dinamo, pero no cabe excluir la posibilidad de que exista también un campo magnético primordial no detectado de hasta un millón de gauss de intensidad aprisionado en el núcleo estable del Sol.

La hipótesis de los campos primordiales adquiere mayor interés en el caso de estrellas con campos mucho más intensos que el solar. Una clase de estrellas con campos intensos es la llamada de estrellas magnéticas A. (Las estrellas A tienen una temperatura en super-

ficie de unos 10.000 grados Kelvin). Fueron descubiertas a finales del decenio de 1940 por Harold D. Babcock y Horace W. Babcock, de los Observatorios de los Montes Wilson y Palomar, quienes midieron sus campos con un magnetómetro sensible que habían desarrollado. Los campos de estas estrellas varían en intensidad desde unos cientos hasta 34.000 gauss.

Se ha puesto de manifiesto, de entonces acá, que el campo de una estrella magnética *A* es un dipolo que gira más o menos rígidamente con la propia estrella. Además, el eje del campo tiende a ser perpendicular al eje de rotación en vez de correr más o menos paralelo a él, como en los planetas y otras estrellas. Las estrellas tienen, además, una modesta velocidad de rotación (un período de alrededor de un mes); de manera que las estrellas de rotación más rápida poseerán, pues, campos magnéticos más débiles que las que giran más lentamente. La falta de correlación entre la intensidad y el modo del campo, entre la intensidad del campo y la velocidad de rotación y entre la orientación del campo y el eje de rotación hace improbable que los campos de esas estrellas se generen a través del mecanismo de dinamo.

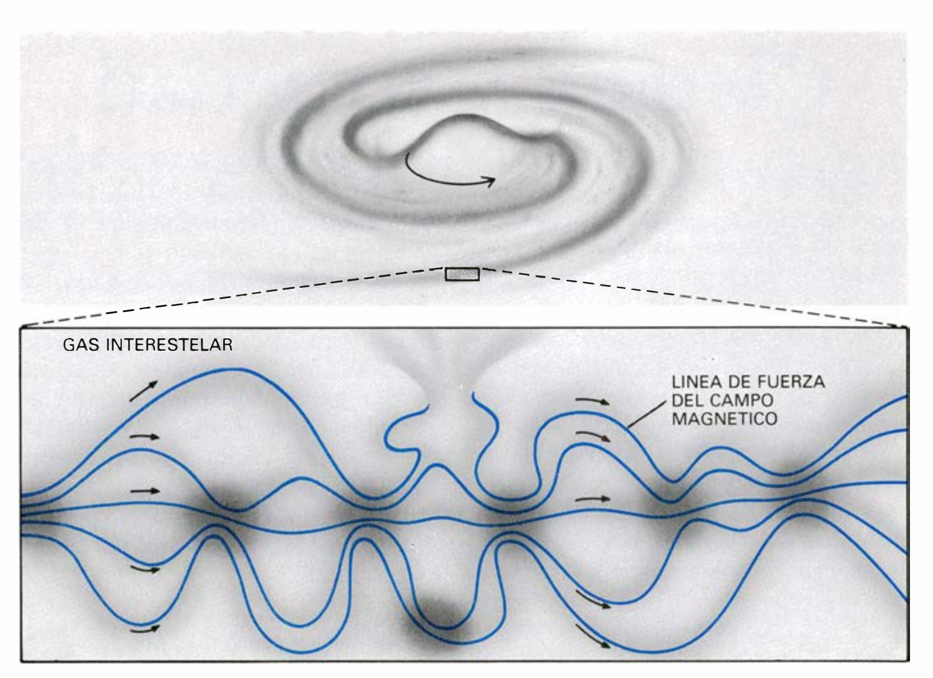
Se ha sugerido también que la primordialidad del campo magnético galáctico se basaría en que el tiempo de debilitamiento resistivo de un cuerpo tan grande y tan conductor como la galaxia es mayor que la edad de ésta. El efecto de los electrones circulando a lo largo de las líneas del campo galáctico en el espacio interestelar sobre las radioondas planas polarizadas procedentes de fuentes lejanas suministra una medida de la intensidad del campo. Observaciones de la variación del plano de polarización con la frecuencia (efecto de rotación de Faraday) muestran que la intensidad del campo está comprendida entre 2×10^{-6} y 3×10^{-6} gauss. Todo lo que se necesitaría para



8. COMIENZO DEL CICLO MAGNETICO del Sol con dos bandas de campo azimutal de signo opuesto en cada hemisferio (*1*). La banda a latitudes altas aparece como las regiones magnéticas bipolares "efímeras" visibles en los magnetogramas del Sol. La banda más baja aparece en forma de manchas solares emparejadas; es decir: puntos de entrada y salida de un campo azimutal que se eleva por encima de la superficie de la zona convectiva. El campo meridiano generado por convección en cada banda de campo azimutal es estirado por el gas en giro no uniforme de la zona convectiva para formar un nuevo campo azimutal (*2*). El nuevo campo refuerza al original en el borde meridional de las bandas y lo anula en el borde septentrional (*3*). Como resultado, las bandas de campo azimutal emigran hacia el ecuador, mientras a latitudes altas se crea una nueva banda de signo opuesto. En el ecuador, las bandas de campo azimutal son anuladas por bandas de signo contrario procedentes del otro hemisferio (*4*). Las manchas solares asociadas con cada banda de campo decrecen en número e intensidad a medida que la banda se debilita lentamente (*5*). La recuperación desde el mínimo de número de manchas solares hasta el máximo siguiente es bastante rápida (*6*). La serie dibuja un cambio de polaridad: la mitad del ciclo magnético solar de unos 22 años.

dar cuenta del campo actual sería, quizá, que el gas difuso a partir del cual se formó la galaxia hubiera sido dirigido por un campo de unos 10^{-9} gauss. Este campo se comprimiría cuando el gas colapsara para formar la galaxia. Se ha argumentado que la compresión y subsiguiente alargamiento de las líneas de fuerza en la galaxia en rotación no uniforme podría haber producido un campo de la intensidad observada. Sin embargo, un trabajo reciente de Jacques P. Vallée, del Consejo Nacional de Investigación del Canadá, establece un límite superior de 10^{-11} gauss para el presente campo intergaláctico, intensidad menor que los 10^{-9} gauss supuestos en los razonamientos a favor del origen primordial del campo galáctico.

Lo que es más importante: el sistema compuesto por el gas interestelar, los rayos cósmicos y el campo magnético galáctico es altamente inestable. El gas tiende a gravitar hacia el plano central del disco galáctico; el campo magnético tiende a ondularse hacia arriba, alejándose del plano. Aunque no hay información suficiente para elaborar un modelo cuantitativo de la pérdida turbulenta de campo, los cálculos posibles indican que las vigorosas ondulaciones en pequeña escala del campo hacen que éste se bambolee alejándose del gas del disco a una velocidad suficiente para impedir que un campo primordial haya persistido hasta la fecha. La pérdida turbulenta reduce el tiempo de debilitamiento del campo a unos 300 millones de años, en tanto que la edad de la galaxia es de 10.000 millones de años.



9. CAMPO MAGNETICO GALACTICO, cuya dirección es predominantemente azimutal (paralelo al plano de la galaxia), aunque la dirección del campo varía ampliamente. La dirección estaba originariamente determinada por la polarización de la luz procedente de estrellas distantes. Los granos de polvo en el gas interestelar entran en rotación tras los choques provocados entre sí o con átomos de hidrógeno. Cada grano de polvo lleva una carga eléctrica; tiende, así, a alinearse con su eje mayor perpendicular a una línea del campo magnético, alrededor de la cual gira. La polarización de la luz absorbida y reemitida por el polvo interpuesto entre una estrella y el sistema solar está determinada, pues, por la orientación del campo magnético galáctico. Las intensas fluctuaciones locales del campo magnético vienen producidas por su interacción con el gas interestelar y los rayos cósmicos a los que confina y por los que es amplificado. Se cree que las explosiones de supernova de las estrellas de gran masa arrojan el gas interestelar y los rayos cósmicos hacia fuera del plano de la galaxia, creando las capas esféricas de gas visibles en los mapas de radioemisión del hidrógeno atómico, uno de los componentes del gas. El campo magnético delineado en el interior del gas es transportado con él y comprimido en la periferia de la capa. Si este campo no resulta ser suficientemente intenso para equilibrar la presión del gas, dirigida hacia fuera, el gas y el campo magnético inmerso en él salen entonces barridos hacia el espacio. Tal turbulencia contribuye apreciablemente a la velocidad de disipación del campo galáctico. La velocidad de pérdida hace improbable un campo primordial y apoya la hipótesis de que sea regenerado por gas en rotación y convección.

El mecanismo de dinamo explica con éxito lo que se sabe acerca del campo galáctico. Las ecuaciones de dinamo, vueltas a formular en la forma apropiada para un delgado disco con la conductividad eficaz del gas interestelar turbulento, muestran que la galaxia podría regenerar un campo magnético a una velocidad que compensara la pérdida por escape turbulento y debilitamiento resistivo. Se sabe que el disco de la galaxia gira con una velocidad angular que decrece desde el centro hacia la periferia. Las explosiones de supernova y la brusca formación y comienzo del brillo de estrellas luminosas de gran masa producen violenta turbulencia o convección. A causa de la rotación, la turbulencia es ciclónica. El campo primario de la glaxia es azimutal, extendiéndose alrededor del disco de la galaxia. Las componentes meridianas del campo galáctico no se pueden observar porque se pierden en las intensas fluctuaciones locales del campo azimutal.

Hay una última cuestión que merece consideración, aun cuando, a causa de la falta de pruebas, es más filosófica que científica. El mecanismo de dinamo no suministra, desde un comienzo, la creación de un campo magnético, sino solamente la reproducción, esto es, la intensificación de campos preexistentes. ¿Dónde se originaron entonces los campos que la dinamo intensifica? Puede no ser necesario recurrir a la idea de un campo primordial, cuya existencia debiera a su vez explicarse. En 1941, Ludwig F. Biermann, de la Universidad Humboldt de Berlín, indicó que existen efectos de "batería atómica" en el gas ionizado móvil de las estrellas. Por ejemplo, las isotermas, o superficies de igual temperatura en una estrella en rotación, no coinciden con las superficies elipsoidales de nivel en las que se equilibran los potenciales centrífugo y gravitatorio. Como resultado de las fuerzas no compensadas, los electrones libres deslizan a lo largo de esas superficies, creando débiles corrientes meridianas. Si una estrella estuviera inicialmente libre de campo magnético, el efecto de batería de Biermann llegaría a desarrollar, a lo largo de un período de tiempo suficiente, un débil campo azimutal. El efecto de batería no puede dar cuenta, por sí solo, de campos de la intensidad de los observados hoy, pero garantiza que, si todo lo demás fallase, las estrellas y las galaxias estén sembradas de débiles campos magnéticos sobre los que podría actuar el potente mecanismo hidromagnético. Efectos similares, debidos a ligera radiactividad o a heterogeneidad térmica, sembrarían los planetas de débiles campos magnéticos.

Cuando se consideran todas las pruebas existentes, parece como si los campos magnéticos se produjeran en virtud de movimientos fluidos en todas las escalas, desde los 2000 kilómetros del núcleo de metal líquido de un planeta hasta el millón de kilómetros de una estrella o los 10^{18} kilómetros de una galaxia. Queda mucho por conocer, y tanto sobre dinamos cuanto sobre cuerpos singulares cuyos campos son demasiado intensos para poderse explicar por ese efecto conocido. Pero la existencia de problemas intrigantes aviva el interés del tema.

Ciencia y sociedad

Farmacopea elemental

Las farmacopeas de los países industrializados cuentan con unas 1500 drogas. Todas tienen sus aplicaciones, pero sólo unas pocas podrían catalogarse entre las básicas para la salud de la población humana. ¿Cuáles son? Se ha avivado la cuestión a raíz de la situación de precariedad en que se encuentran muchos países en vías de desarrollo, con un sistema sanitario rudimentario, en manos de las compañías farmacéuticas extranjeras y carente de transporte y almacenamiento idóneos para el buen estado de los medicamentos. El Comité de Expertos de la Organización Mundial de la Salud sobre el Uso de Drogas Fundamentales ha abordado el problema y ha publicado una lista de 249 sustancias que se adaptarían a esas condiciones; de esa lista principal, 22 productos se consideran esenciales para una asistencia sanitaria mínima.

En su elección, el comité aplicó varios criterios. Los fármacos deben satisfacer las necesidades sanitarias de la mayoría de la población de un país, y por tanto "los habrá en todo momento, en cantidades adecuadas y en las dosis convenientes". La elección de fármacos que efectúe un país dependerá de muchos factores: enfermedades predominantes, instalaciones asistenciales y conocimientos y experiencia del personal sanitario. Más aún, "sólo deberían elegirse los medicamentos de los que existen datos favorables por su eficacia y seguridad, recogidos de estudios clínicos apropiados, y de los que se ha obtenido pruebas de sus resultados en usos generales en diversos estados médicos".

Guiado por estos criterios, el comité compuso la lista principal, dividida en 27 categorías, entre las que se cuentan: analgésicos; antídotos; drogas antiinfecciosas; productos hematológicos y medicamentos que afectan a la sangre; fármacos para trastornos cardiovasculares, dermatológicos y gastrointestinales; hormonas; relajantes musculares; preparaciones oftalmológicas; drogas psicoterapéuticas; fármacos del tracto respiratorio y vitaminas y minerales.

Las 22 sustancias incluidas en la "lista modelo de fármacos para la asistencia sanitaria primaria" por el comité son: ácido acetilsalicílico (aspirina), carbón activo, un antiácido, un fármaco antihemorroidal, atropina (un antiespasmódico), ácido benzoico con ácido salicílico, benzoato de bencilo, loción de calamina, clorhexidina en solución, cloroquina, clorfenamina, efedrina (para el asma), ergometrina (para las hemorragias perinatales), yodo, ipecacuana, hierro y ácido fólico (suplemento nutritivo para el embarazo), gammahexano, mebendazol, sales orales de rehidratación, paracetamol y ungüento ocular de piperacina y tetraciclina. El comité afirma que los medicamentos de la relación "pueden administrarse eficazmente y sin peligro por personas responsables con exigua formación médica".

Esperando a Halley

La aparición del cometa Halley, en 1910, fue tan espectacular que, fotografiada por los mayores telescopios del momento, se convirtió en una especie de alegoría astronómica. Poco después de que alcanzara su distancia más próxima al Sol, presentaba una brillante coma (la nube de gas y polvo que envuelve al núcleo) y una luminosa cola doble que se extendía casi por un cuarto del cielo. ¿Qué aspecto mostrará el cometa, a su regreso, en 1985 y 1986? Ian Halliday, del Instituto Herzberg de Astrofísica de Ottawa, analiza la situación en un trabajo que publica *The Journal of the Royal Astronomical Society of Canada*. Para el aficionado, las perspectivas no son muy halagüeñas.

El cometa se ha avistado ya con el telescopio de 200 pulgadas del Monte Palomar. Está, pues, en viaje de vuelta. (Algunos cometas desaparecen a causa de perturbaciones orbitales provocadas por los planetas mayores.) Según Halliday, la geometría de la órbita del cometa será la menos favorable de las 29 apariciones registradas. Sigue una órbita elíptica alargada, inclinada en un ángulo de 18 grados con respecto al plano de la eclíptica (el plano principal de las órbitas de los planetas). Un cometa brilla al máximo y exhibe su más alto grado de actividad cerca del perihelio (el punto de su órbita más cercano al Sol). Cuando el cometa de Halley se aproxime a su perihelio, el 9 de febrero de 1986, la Tierra se hallará en la otra cara del Sol, hasta el punto de que cometa, Sol y Tierra casi se alinearán. Desde ésta, el cometa aparecerá menos de cinco grados al norte del Sol y algo hacia el oeste. Saldrá, pues, una hora antes que el Sol, en una región de fuerte luz crepuscular.

Los acercamientos máximos del cometa a la Tierra tendrán lugar el 27 de noviembre de 1985, poco después de que atraviese el plano de la eclíptica en su viaje al Sol, y el 11 de abril de 1986, alrededor de un mes después de traspasar de nuevo el plano de la eclíptica, en su salida del centro del sistema solar. En ninguno de esos casos se hallará el cometa muy próximo a la Tierra; y en ambas ocasiones se encontrará a más de una unidad astronómica del Sol (la distancia que separa la Tierra del Sol). A esa distancia no es probable que el cometa sea muy activo.

A pesar de las dificultades que plantea su órbita, los astrónomos esperan el paso del cometa con gran impaciencia. La razón estriba en que el cometa de Halley es el único de los de período corto (inferior a los 200 años) que exhibe, de forma fidedigna, una amplia gama de actividades cometarias, y cuya órbita se ha determinado al detalle. Ello permite planificar con antelación las redes de instrumentos y las misiones espaciales de interceptación que puedan tomar las imágenes y los datos que compensen el paso poco favorable. La International Halley Watch (Observación Internacional de Halley), un comité auspiciado por la Unión Astronómica Internacional, está organizando ya para la ocasión redes instrumentales fotométricas, espectroscópicas y de otros tipos. Se han previsto también tres misiones espaciales de interceptación: la astronave *Giotto*, construida por la Agencia Espacial Europea, la *Venera-Halley*, de la Unión Soviética y otros países europeos, y la *Planet A*, construida por Japón.

En la actualidad, el cometa se reduce a un núcleo de hielo, apreciable al telescopio porque refleja la luz solar. Sin embargo, conforme se aproxime al Sol, el hielo comenzará a sublimarse. El gas en expulsión, junto con las partículas que transporte, crearán la coma, que ganará en luminosidad a medida que las partículas de polvo dispersen la luz y los fragmentos de moléculas disociadas por la radiación solar emitan fluorescencia. En la coma se generarán chorros o abanicos cuando reciba la intensa radiación del Sol. Por último, la presión de la luz solar expulsará partículas de polvo de la coma, creando una cola de polvo, larga y curvada. El viento solar, las partículas con carga eléctrica que se proyectan del Sol, ionizará el gas de la coma. El gas ionizado se verá arrastrado de vuelta por los campos magnéticos que genere el viento solar, lo que provocará la aparición de una cola recta, en la que probablemente se formen corrientes filamentosas, nudos y otras estructuras. El conocimiento de

estas actividades revelaría importante información sobre el cometa. Por ejemplo, la observación de los chorros y abanicos de la coma permitiría la deducción de la orientación y velocidad de rotación del núcleo.

Las redes de instrumentos terrestres tienen, además, otro objetivo: seguir al cometa durante y después del perihelio, la porción menos previsible de su órbita. En su mayor parte, las órbitas de los cometas pueden predecirse a partir de la ley de la gravedad de Newton. Edmund Halley advirtió que los grandes cometas avistados en 1531, 1607 y 1682 habían seguido, en el cielo, trayectorias casi idénticas, por lo que debía tratarse de un mismo objeto que recorriera una órbita fuertemente elíptica alrededor del Sol. En 1705 predijo, a partir de sus cálculos de la órbita del cometa, que éste reaparecería en 1758. Algún tiempo después advirtió la conveniencia de tomar en consideración la atracción gravitatoria de Júpiter y Saturno (no se habían descubierto aún Urano ni Neptuno), junto con la del Sol, por lo que revisó su predicción para el perihelio, situándolo a finales de 1758 y principios de 1759. Al confirmarse la predicción, se dio al cometa su nombre.

Cuando un cometa se aproxima al Sol, intervienen fuerzas no gravitatorias menos predecibles, que podrían modificar la órbita del cometa de Halley lo bastante para que las misiones espaciales erraran su objetivo. La expulsión de gases que provoca la radiación solar es más vigorosa en la cara crepuscular, caliente, del cometa en rotación, que en la cara matutina, más fría. El cometa de Halley gira en el sentido de su avance alrededor del Sol, por lo que el empuje que crea la expulsión de gases, comparable al de los cohetes, presenta una componente hacia delante que aleja del Sol al cometa, incrementando su período orbital en varios días. Por tanto, la información que aporten las redes terrestres es esencial para que las misiones espaciales proyectadas para marzo de 1986, cuando el cometa de Halley haya superado el perihelio y se aproxime al plano de la eclíptica, logren su objetivo.

Menos lazos

No cabe duda de que en los países industrializados la institución matrimonial está cambiando. ¿Cuál es la tendencia? En un artículo que publica el *Bulletin of the American Academy of Arts and Sciences*, Kingsley Davis, de la Universidad de California del Sur, examina datos demográficos sobre los matrimonios celebrados en los países más prósperos desde la segunda guerra mundial, y concluye que el matrimonio está perdiendo su conexión con sus dos funciones principales: la intimidad familiar y la procreación.

Desde 1945 hasta 1960, el número de matrimonios fue en aumento en los países industrializados. Se constató en general un descenso de la edad en que hombres y mujeres se casaban por primera vez y un aumento global de la población casada. A partir de la década de 1960, y sobre todo en los 70, las tendencias se invirtieron. En 1960, el 14 por ciento de las mujeres estadounidenses con una edad de 30 a 34 años estaban solteras; en 1980, la fracción era del 27 por ciento. Otras naciones industrializadas muestran porcentajes similares de mujeres solteras. El cambio se debe al aumento de la frecuencia de divorcios y a la tendencia de los jóvenes a retrasar la boda. Tras un descenso constante a partir de 1945, la edad del primer matrimonio comenzó a aumentar a partir de 1970.

El cambio de hábito social ha producido un gran grupo de jóvenes solteros. No quiere ello decir que ese grupo rechace los placeres de la convivencia. Desde 1977 hasta 1980, el número de parejas casadas no varió mucho en los Estados Unidos, pero el número de parejas no casadas aumentó en un 63 por ciento. En Noruega, en 1977, el 43 por ciento de las jóvenes de 18 a 19 años que convivían con un hombre eran solteras. La proporción de mujeres que viven en tales "uniones consensuadas" decrece con la edad, pero, entre las jóvenes, las uniones informales compiten vigorosamente con las formales.

Sugieren algunos que las uniones sin formalizar no son más que matrimonios exentos de ciertas trampas legales más o menos triviales, aunque Davis señala que existen entre ambos dos diferencias significativas. Las uniones informales corresponden a gente joven, son menos estables y producen menos hijos que las matrimoniales. En cualquier caso, la proporción de población que convive con otra persona, sea en matrimonio o no, está descendiendo; las uniones consensuadas no han compensado el descenso de la tasa matrimonial.

El matrimonio, además de sufrir la competencia de otros tipos de relaciones en la procuración de intimidad doméstica, pierde rápidamente también el monopolio de la reproducción y la educación de los hijos. Las parejas casadas son cada vez menos fértiles. De 1945 a 1949, la tasa neta de producción en 19 de los países de mayor desarrollo económico era del 1,27. (La tasa neta de reproducción indica si la población se sostiene; el nivel de sustitución es 1.) De 1975 a 1979, la tasa fue, en esos 19 países, del 0,87, muy por debajo del nivel de sustitución. En países como los Estados Unidos, Gran Bretaña y Suecia, la tasa hubiera sido bastante inferior de no ser por la inmigración procedente de países pobres, cuya fertilidad es relativamente elevada.

La reducción de la fertilidad es, de hecho, mucho mayor de lo que sugiere el cambio de la tasa de reproducción. La razón es que crece el número de madres solteras. En 1950, la proporción de mujeres blancas con hijos era, en los Estados Unidos, de una soltera por cada 17 casadas; en 1970 era de una por cada cuatro. Entre las mujeres negras norteamericanas, la probabilidad de tener un hijo es casi la misma para las solteras que para las casadas. La fertilidad de las uniones consensuadas es bastante baja, por lo que la mayoría de los nacimientos de mujeres solteras corresponden a las que viven solas.

Si se suman los hijos de padres divorciados a los de padres solteros, resulta que una gran porción de los niños de los países desarrollados se están criando en un ambiente que no es el del matrimonio tradicional. En 1970, un 15 por ciento de los niños norteamericanos no convivía con sus progenitores; en 1981, la proporción era ya del 24 por ciento. Cada vez son menos los niños que crecen en un matrimonio no disuelto. Y a la inversa, cada vez son menos los matrimonios que tienen hijos: en 1981, el 49 por ciento de los matrimonios estadounidenses no convivía con hijos propios de menos de 18 años de edad.

Los datos recogidos por Davis muestran que el matrimonio se aparta cada vez más de sus funciones tradicionales. Las relaciones conyugales son más frecuentes en una unión informal y la reproducción es más frecuente fuera del matrimonio; más aún, fuera de cualquier unión. Para justificar los cambios observados en la institución matrimonial y, en particular, el descenso de la fertilidad de las parejas casadas, Davis señala varios factores. Uno de los más significativos es la probabilidad, cada vez mayor, de que la mujer casada trabaje fuera del hogar. Un 60 por ciento de las mujeres casadas trabajan, en los Estados Unidos, fuera de casa, proporción sólo ligeramente inferior a la de las mujeres solteras, que es del 65 por ciento. A falta de las condiciones necesarias para el cuidado de los hijos, cuanto mayor es la proporción de mujeres con trabajo, menor es la tasa de fertilidad.

Materia interestelar en meteoritos

Las condritas carbonáceas, meteoritos primitivos, incorporan materia originada fuera del sistema solar, como la expelida por las supernovas y otras estrellas

Roy S. Lewis y Edward Anders

El sistema solar es mucho más joven que el universo; cuenta con 4500 millones de años de existencia frente a los 10.000 o 15.000 millones de años de éste. Hubo de formarse, pues, a partir de materia más antigua que tenía su propia historia. Hasta 1969 no se había encontrado restos, ni en meteoritos ni en planetas. Al proponer los teóricos un origen caliente para el sistema solar, que habría provocado la vaporización de todos los sólidos preexistentes, se infería que el sistema solar arrancó con una lista limpia.

¿Cómo podría reconocerse la materia presolar? Tal vez por su edad, y, mejor, por su composición isotópica. Las estrellas reconstruyen continuamente los elementos químicos, para arrojarlos de nuevo al espacio interestelar, donde, con el tiempo, forman la nueva generación de estrellas. La composición elemental y la composición isotópica de la materia eyectada varía de una estrella a otra, de acuerdo con la temperatura, masa y estado de evolución de la estrella. La composición isotópica es un sello distintivo de larga duración, ya que pocos procesos pueden alterarla (fuera de las reacciones nucleares) y, aun entonces, en cuantía limitada y previsible. Por tanto, hay que ir a buscar materia de composición isotópica poco corriente que no se justifique mediante procesos conocidos o plausibles en el sistema solar.

El primer indicio de materia presolar llegó en 1969 cuando David C. Black y Robert O. Pepin, de la Universidad de Minnesota, estudiaron la composición isotópica del gas noble (químicamente inerte) neón en las condritas carbonáceas: piedras de un color gris oscuro, de aspecto poco llamativo, que son los meteoritos más primitivos que se conocen. Escondida entre grandes cantidades de neón normal, había una porción pequeña del mismo muy enriquecida en el isótopo neón 22. Dada la rareza y escasez de los gases nobles, no se podía descartar un origen local. En 1973, Robert N. Clayton, Lawrence Grossman y Toshiko Mayeda, de la Universidad de Chicago, hallaron que ciertos minerales de la condrita carbonácea de Allende estaban enriquecidos en el isótopo oxígeno 16 en una cuantía del 5 por ciento con respecto al oxígeno normal, por lo menos. Después de descartar un origen local, arribaron a la conclusión de que parte del oxígeno fabricado por reacciones nucleares en estrellas distintas del Sol había llegado, de alguna manera, al meteorito.

A éste siguieron otros descubrimientos. Hoy se cuentan hasta 19 los elementos químicos de las condritas carbonáceas que presentan anomalías isotópicas. No se trata, por supuesto, de una lista cerrada. Parte de la materia presolar sobrevivió en las regiones más frescas del primitivo sistema solar. El problema es hallar esos trozos de materia presolar y descifrar su historia de nucleosíntesis estelar y química interestelar.

El carbono ha proporcionado una cosecha particularmente rica de componentes presolares. Cuatro se llevan contados ya. Perfectamente agazapados, les delató el hallarse "marcados" con gases nobles o hidrógeno anómalos. Seguiremos las cuatro pistas del descubrimiento, una vez descritos los meteoritos donde se observan las anomalías.

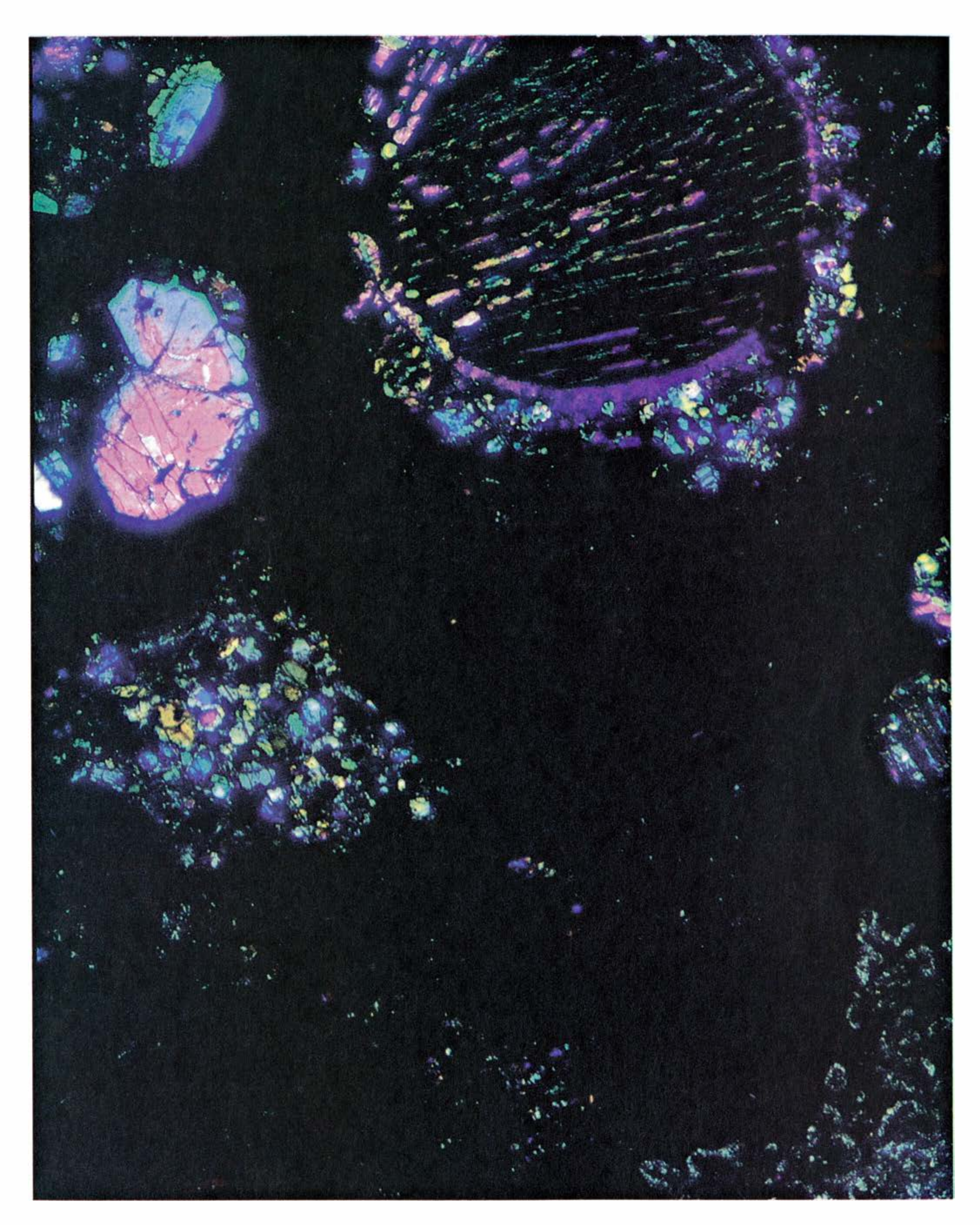
Condritas carbonáceas

Se conocen unas 40 condritas carbonáceas, que representan aproximadamente el 4 por ciento de todos los meteoritos que caen. Se dividen en tres clases principales, que se designan por C1, C2 y C3 según su constitución mineralógica y contenido decreciente de cuerpos volátiles, tales como agua, carbono y nitrógeno. Su temperatura de formación (temperatura a la que cesaron de reaccionar con el gas de la nebulosa solar o nube a partir de la cual se condensó el sistema solar) aumenta desde C1 hasta C3; si la temperatura subía hasta el centro de la nebulosa, las C1 provendrían de la parte más externa. (Se formaron a temperaturas de unos 360 grados Kelvin. A otros 60 grados más surgieron las C3.) Todas las condritas carbonáceas sufrieron alteraciones en sus cuerpos padres, asteroides seguramente. Las C1 y C2 estuvieron allí expuestas al agua líquida; las C3 sufrieron recalentamientos de hasta 600 grados Kelvin.

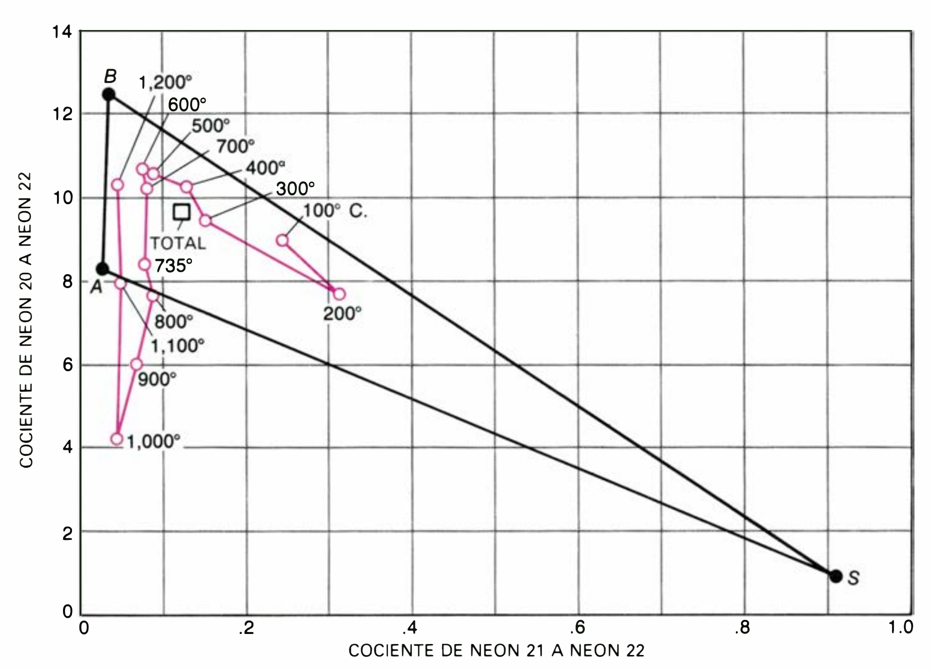
La textura de las condritas está consituida por una matriz de grano fino, que encierra partículas de grano más grueso. Las partículas de la matriz, ricas en materias volátiles, tienen tamaños de 10^{-4} a 10^{-6} centímetros; constan de silicatos y materia carbonácea. El tamaño de las partículas mayores varía de 0,01 centímetros a 1 centímetro. Son pobres en materias volátiles. Algunas de ellas son cóndrulos, o partículas redondeadas parecidas a perdigones. Otras, aunque irregulares, están compuestas de los mismos minerales: olivino [$(Mg, Fe)_2 SiO_4$], piroxeno [$(Mg, Fe)SiO_3$], troilita (FeS) y níquel-hierro. Todavía hay otras que son inclusiones refractarias (resistentes al fuego) de minerales ricos en calcio, aluminio y titanio, pero pobres en silicio. Visto desde hoy, no es ninguna sorpresa que las anomalías isotópicas se encuentren en la matriz, que se ha calentado muy poco, y en las inclusiones refractarias, altamente resistentes al calor. Las anomalías que explicaremos a continuación se encuentran principalmente en la matriz.

Neón *E*

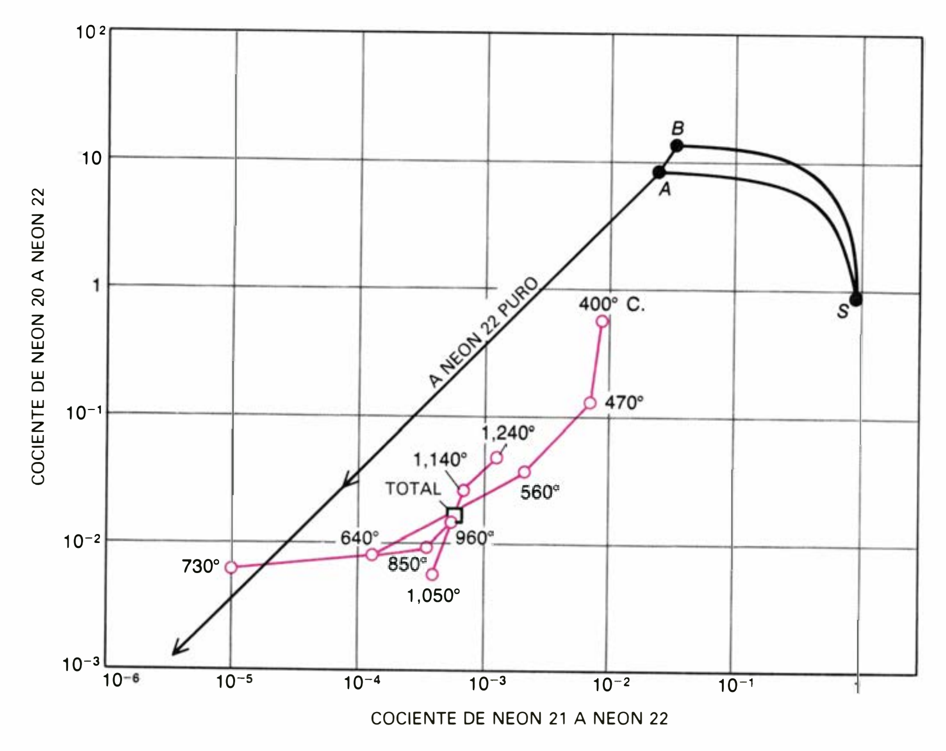
Comenzaremos con los gases nobles. Destacan entre los elementos hallados



1. CORTE FINO DEL METEORITO DE ALLENDE fotografiado bajo luz polarizada. Revela la contextura característica de las condritas carbonáceas. El campo de visión tiene aproximadamente un milímetro de anchura. En la parte superior derecha hay un cóndrulo o gotita de silicato solidificado. En el resto, pedazos irregulares de silicatos, sulfuros y ferroníquel. En la parte inferior derecha destaca una inclusión refractaria (de alto punto de fusión) del mineral olivino. El resto del campo, negro en la fotografía, es una matriz de grano fino que consta de silicatos y materia carbonácea. Hasta ahora se han encontrado en las condritas carbonáceas proporciones de isótopos distintas de las que muestra la materia del sistema solar en 19 elementos químicos, la mayor parte en las inclusiones y en la matriz; la lista no está cerrada. Fotografía tomada por Glenn J. MacPherson, de la Universidad de Chicago.



2. DIAGRAMA DE TRES ISOTOPOS que revela el neón anómalo en el meteorito de Orgueil. Se ha calentado una muestra del meteorito hasta temperaturas progresivamente crecientes y se ha analizado la fracción de gas desprendido a cada temperatura para determinar su proporción en los isótopos neón 20, 21 y 22. Las dos primeras fracciones se inscriben en el área triangular que representa mezclas de las tres clases de neón que son de esperar en los meteoritos: el primario (*A*), que procede de la nebulosa solar o nube gaseosa de la que se formó el sistema solar; el neón solar (*B*), introducido en el meteorito por el viento de iones procedente del Sol, y el cosmogénico (*S*), producido por choques entre los rayos cósmicos y los núcleos atómicos dentro del meteorito. Hay además una sucesión de fracciones que se aparta del triángulo. Algo encerrado en el meteorito de Orgueil desprende un tipo de neón altamente enriquecido en neón 22. Se le llamó después neón *E*, reservadas ya las letras *C* y *D* para sendos componentes menos extremos. El experimento fue realizado, en 1969, por David C. Black y Robert O. Pepin.



3. EN UN EXPERIMENTO POSTERIOR, M. H. A. Jungck y Peter Eberhardt, de la Universidad de Berna, aplicaron el calentamiento escalonado y la representación en diagrama de tres isótopos a una pequeña muestra de mineral grandemente enriquecido en neón *E*, que prepararon a partir del meteorito de Orgueil. El neón *E* mostró ser esencialmente neón 22 puro (más del 99 por ciento de las fracciones desprendidas desde 640 hasta 1050 grados Celsius). Por lo que se deduce, el neón *E* se formó en la desintegración radiactiva del sodio 22 en granos alrededor de una estrella, nova quizá. Cuando se condensó el sistema solar, los granos quedaron incorporados dentro del meteorito. El diagrama es logarítmico.

en los meteoritos. A pesar de ser altamente volátiles y químicamente inertes, no se condensaron ni siquiera en los más primitivos meteoritos; de ahí que se presenten en una mínima fracción de su proporción en el Sol, variando desde 10^{-5} para el xenón hasta 10^{-9} para el helio y el neón. Sin embargo, esta pequeña cantidad de gas se encuentra estrechamente aprisionada en el meteorito, liberándose sólo a temperaturas muy altas, cuando el mineral anfitrión comienza a fundirse o descomponerse.

Black y Pepin trataron de desenmarañar tres clases de neón que se sabe están presentes en los meteoritos primitivos: el neón primitivo o planetario, llamado también neón *A*, que quedó aprisionado a partir de la nebulosa solar; el neón solar, llamado neón *B*, que consta de iones de neón procedentes del viento solar implantados en meteoritos que casualmente se encontraban en la superficie de su cuerpo originario; y el neón cosmogénico, también llamado neón *S*, formado cuando los rayos cósmicos que atraviesan el meteorito fracturan, o rompen, núcleos atómicos que había en su camino. Cada tipo tiene diferentes proporciones de las tres clases de neón.

Black y Pepin emplearon la técnica del calentamiento escalonado, por la que se calienta una muestra del meteorito hasta temperaturas progresivamente más altas y el gas liberado se analiza en un espectrómetro de masas. Con suerte, los diferentes gases componentes emergen de uno en uno, a medida que se funden sus minerales anfitriones, o se descomponen o hacen permeables. Más frecuentemente, los gases emergen en forma de mezcla, que después ha de separarse de acuerdo con su composición isotópica. La forma más fácil de hacerlo consiste en inscribir dos fracciones isotópicas una respecto a la otra, ambas con el mismo denominador. En el caso del neón, se representa la relación de neón 20 a neón 22 en función de la relación entre neón 21 y neón 22, para cada escalón de temperatura. En tal representación de "tres isótopos", todas las mezclas de dos tipos de neón están sobre una línea recta que las une, las mezclas de tres tipos están en el interior de un triángulo cuyos vértices son los tres, y así sucesivamente.

Las muestras de neón meteorítico medidas hasta el trabajo de Black y Pepin habían caído dentro del triángulo limitado por neón *A*, neón *B* y neón *S*, como era de esperar para mezclas de los tres isótopos. Cuando Black y Pepin analizaron seis condritas C1 y C2 por el

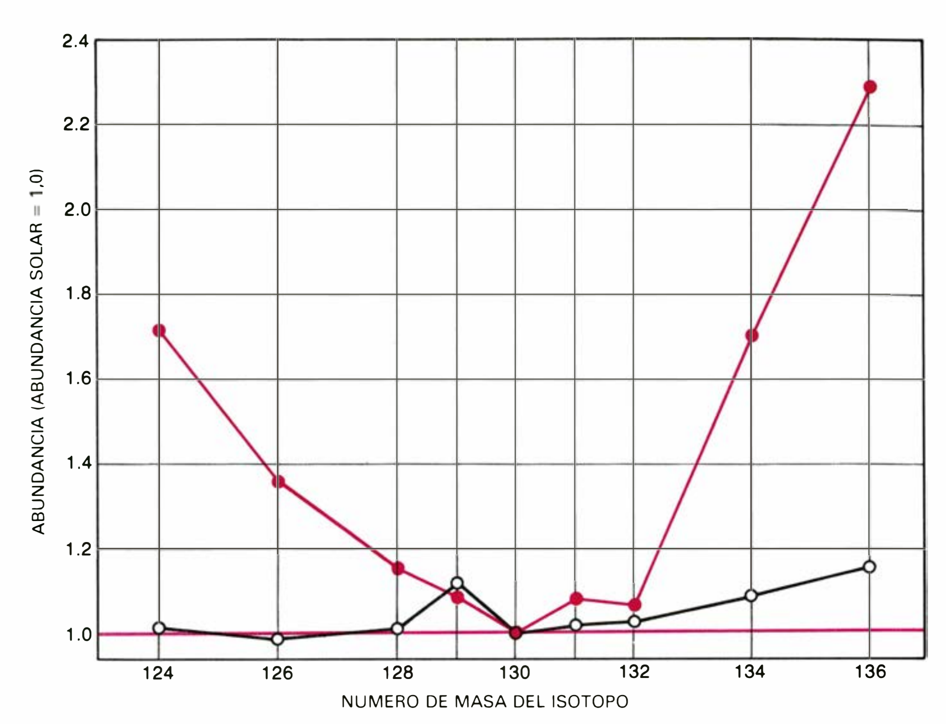
método del calentamiento escalonado, hallaron que las fracciones liberadas entre 800 y 1100 grados Celsius caían sistemáticamente por debajo del triángulo. Existía, pues, un nuevo componente de neón con una relación de neón 20 a neón 22 inferior a 3,4. Le llamaron neón *E*, al haber quedado reservadas las letras *C* y *D* para dos componentes poco importantes de composición menos extrema. Black y Pepin indicaron que el neón *E* no se podía haber formado a partir del neón solar por procesos de fraccionamiento de masa tales como la difusión a través de un sólido o el escape gravitatorio desde un planeta. El depósito inicial de neón hubiera sido absurdamente grande: unas 10^{10} veces mayor que el contenido estimado de neón en la nebulosa solar.

Había otra posibilidad: la intensa irradiación de los granos de polvo por protones (núcleos de hidrógeno) en el primitivo sistema solar. Las reacciones de fractura entre protones y magnesio en los granos podrían producir, además de los isótopos del neón en la proporción del neón *S*, sodio 22. El sodio 22 se desintegraría entonces, con una semivida de 2,6 años, dando neón 22. Donald D. Clayton y sus colaboradores de la Universidad Rice puntualizaron, sin embargo, que la irradiación por protones debería haber producido también excesos de argón 36 y kriptón 80. No se observan tales excesos.

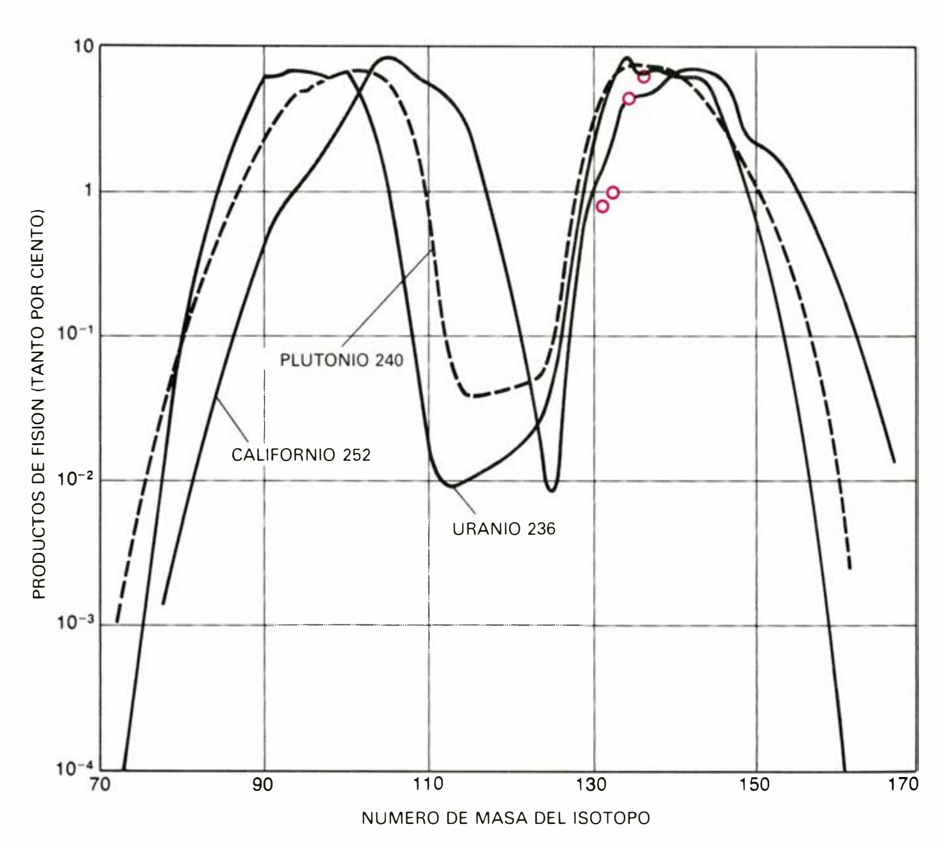
Lo anterior limita nuestra elección a la nucleosíntesis estelar. Las estrellas que han agotado su hidrógeno fabrican neón y elementos próximos a través de reacciones termonucleares en las que toma parte el helio. Así, el helio 4 se combina con oxígeno 16 para producir neón 20. M. Arnould, de la Universidad Libre de Bruselas, y H. Nørgaard, del Instituto Nórdico de Física Atómica Teórica (NORDITA) en Copenhague, han demostrado que el neón *E* podría formarse en estrellas, directa o indirectamente; de un modo directo: bajo condiciones especiales que favorecen grandemente al neón 22 con respecto a los otros isótopos de neón; indirectamente: en forma de sodio 22 que se condensa sobre todo en granos de polvo tras la expulsión de la estrella y se desintegra después para dar neón 22.

Portadores de neón *E*

Había ahí una oportunidad para estudiar el "polvo estelar" y averiguar, quizás, de qué tipo de estrella procedía. El paso siguiente sería, pues, aislar el mineral anfitrión (el "portador") del neón *E* y conocer si el neón *E* muy puro contenía solamente neón 22 o también

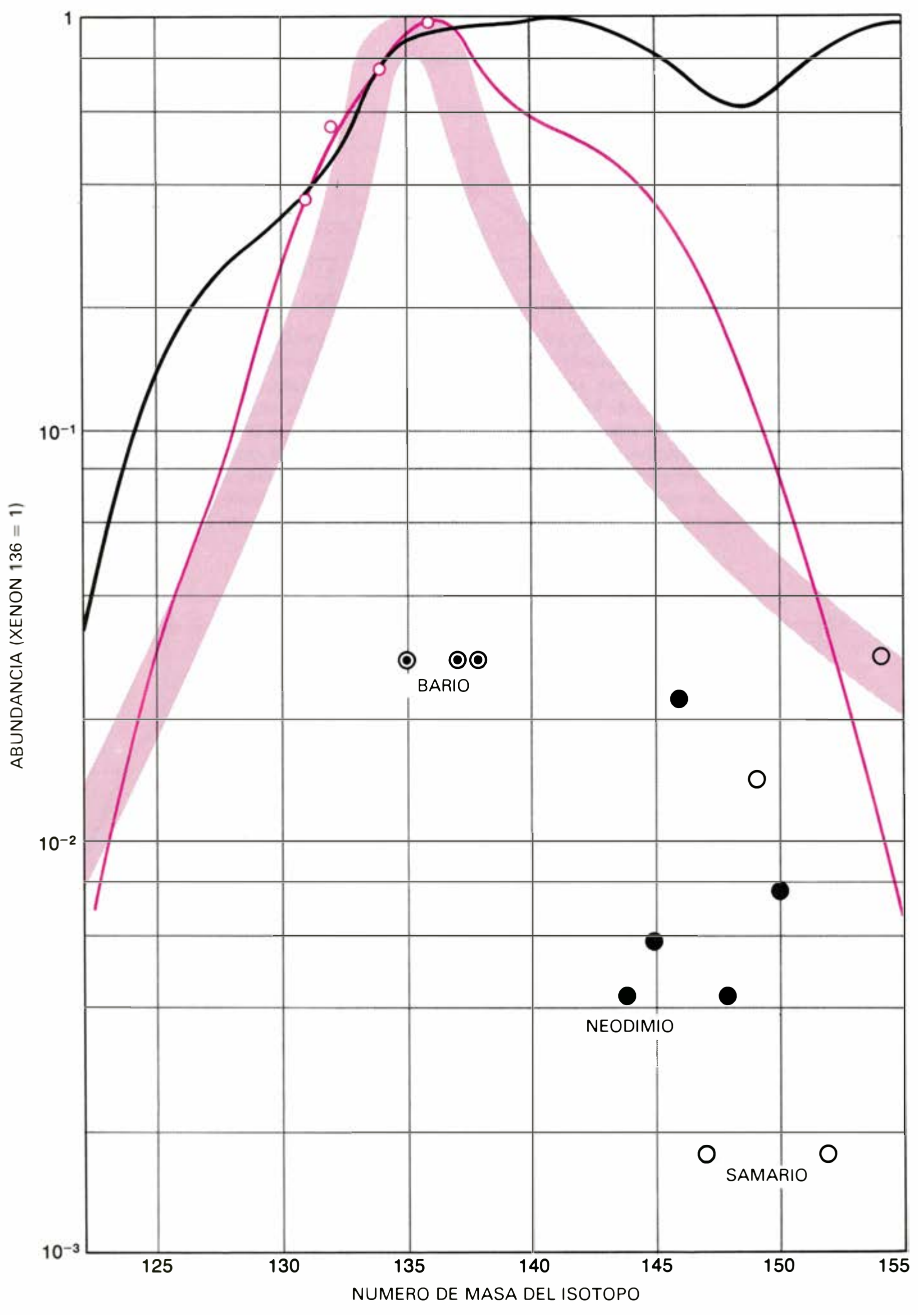


4. XENON ANOMALO, descubierto por John H. Reynolds y Grenville Turner, de la Universidad de California en Berkeley, en el meteorito de Renazzo. En una investigación posterior, B. Srinivasan y los autores de este artículo analizaron, en Chicago, su composición isotópica en el meteorito de Allende. Se examinó primero una muestra de los granos de carbono de Allende; la composición de su xenón (*negro*) difería ligeramente de la correspondiente al xenón solar. Atacaron luego la superficie de los granos con ácido nítrico. El ataque eliminó alrededor del 95 por ciento del xenón. El resto (*en color*) estaba notablemente enriquecido en isótopos pesados y ligeros. El xenón de la muestra era, a ojos vista, una mezcla de xenón primordial de la superficie de los granos y dos clases de xenón anómalo en el interior. Una clase, xenón *H*, estaba enriquecida en isótopos pesados; la otra clase, xenón *L*, en isótopos ligeros.



5. LA FISION DE UN ELEMENTO PESADO en el interior de los meteoritos constituyó la primera explicación propuesta para el origen del xenón *H*. Se sabía que los núcleos de uranio y otros metales radiactivos pesados se rompían asimétricamente, ofreciendo la distribución de productos de fisión con dos máximos que aparece en la ilustración. Los isótopos pesados del xenón descubiertos por Reynolds y Turner (*círculos en color*) siguen esta tendencia general, pero no se ajustan a ninguna curva conocida. Reynolds y Turner sugirieron que el xenón *H* provendría de la fisión de algún elemento superpesado desconocido y ahora extinto en el sistema solar. Pero el xenón *L* debería haberse formado por otra vía.

sus otros dos isótopos. En la resolución del problema intervino de una manera destacada Peter Eberhardt, de la Universidad de Berna, y sus alumnos F. Niederer, M. H. A. Jungck y F. O. Meier. En 1974, Eberhardt, a la sazón en Chicago, deshizo sistemáticamente el meteorito de Orgueil mediante diversos métodos físicos de separación. El mineral magnetita se separaba con un imán. Los silicatos de grano fino se extraían dejándolos formar una suspensión coloidal en soluciones salinas concentradas. En cada paso del proceso, Eberhardt comprobaba si el neón E se había marchado. Halló así que se concentraba en una fracción no magnética de grano grueso. Las separaciones ulteriores mostraron que el neón E tenía al menos dos portadores, uno de baja densidad y temperatura de liberación (menos de 2,2 gramos por centímetro cúbico y 600 grados Celsius) y el otro de densidad y temperatura de liberación más altas (3,5 gramos por centímetro cúbico y 1100 grados). Los correspondientes tipos de neón E se designaron L (de Low, bajo) y H (de High, alto).



6. HIPOTESIS RIVALES sobre el origen del xenón *H*. La hipótesis según la cual surgió por fisión de un elemento superpesado en el interior del meteorito se representa por una curva de productos de fisión calculada para el elemento 114 (*negro*). Una hipótesis rival es la de que el xenón se produjo en una supernova, bien directamente, al capturar neutrones los núcleos ligeros (*banda de color claro*) o indirectamente, al desintegrarse por fisión núcleos superpesados en ella (*línea roja más oscura*). Ambas hipótesis explican los isótopos del xenón *H* (*círculos en color*). Sin embargo, en ambas hipótesis deberían producirse en cantidades similares bario, neodimio y samario. No sucede así; sus límites superiores (*puntos negros*) en un residuo de Allende rico en xenón *H* caen muy por debajo de las curvas teóricas. La hipótesis de la supernova escapa a esta dificultad suponiendo que el bario, neodimio y samario expelidos en una supernova se condensan en granos en una fase en que el xenón, que es más volátil, no podía aún ser aprisionado.

Experimentos posteriores realizados en Berna y Chicago (en los que tomó parte Leo B. Alaerts) establecieron la naturaleza de los portadores. El portador del neón $E(L)$ es carbonáceo. El portador del neón $E(H)$ consta de dos minerales: espinela ($MgAl_2O_4$) y apatito [Ca_2PO_4 (OH, F)]. Mientras tanto, los límites de composición del neón E se iban estrechando de un año a otro, reforzando la posibilidad de que el neón E fuera neón 22 y por tanto derivase del sodio 22. La discusión fue casi resuelta por Jungck y Eberhardt, cuya muestra más pura contenía casi 99 por ciento de neón 22.

Identificados los portadores del neón E, ¿qué puede decirse sobre su origen? La corta semivida, de 2,6 años, del sodio 22, significa que se formó y quedó aprisionado rápidamente, antes de que se desintegrase para originar neón. Donald Clayton ha indicado que estas condiciones se cumplen en las estrellas que explotan: novas y supernovas. Ambas sintetizan elementos explosivamente y los expulsan casi al instante. No deja de tener interés que los primeros minerales que se predijo se condensarían a partir de los gases expulsados por esos dos tipos de estrellas fueran, respectivamente, el carbono y la espinela, los portadores de neón $E(L)$ y neón $E(H)$. Sólo una parte en 10^9 de la masa de cada grano habría de ser de sodio 22 para dar cuenta del neón E que hay hoy. Sin explicación queda, no obstante, el apatito, portador. Nadie sospechaba que esta sustancia se condensara en una explosión estelar, ni siquiera en la nebulosa solar.

Xenón *HL*

Cuando William Ramsay bautizara, en 1898, con el término griego xenón (raro) al elemento que acababa de descubrir, no podía imaginarse con cuánta propiedad le cabría ese nombre al observado en los meteoritos. Tres clases extrañas de xenón, por lo menos, hay en las condritas carbonáceas. Dos de ellas, aunque abundan, son controvertidas; la tercera es rara pero sencilla.

La historia comienza en 1964, cuando John H. Reynolds y Grenville Turner, de la Universidad de California en Berkeley, examinaron la condrita C2 de Renazzo. Iban buscando xenón 129 procedente de la desintegración radiactiva del iodo 129, isótopo descubierto

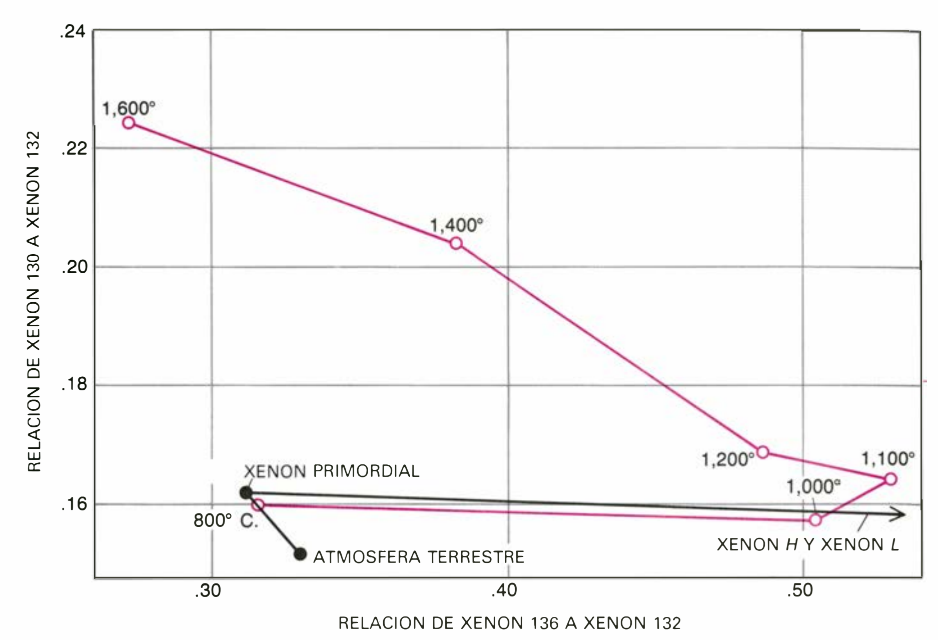
por Reynolds cuya semivida es de 16 millones de años, y que está extinto en el sistema solar. (A lo largo de los 4500 millones de años de la historia del sistema solar, se ha desintegrado tras 280 semividas; se ha reducido, pues, en un factor de 10^{85}.) Para librarse de las grandes cantidades de xenón planetario normal, emplearon la técnica del calentamiento escalonado.

Y hallaron el neón 129. Pero observaron, además, que, en las fracciones liberadas entre 600 y 1100 grados Celsius, los isótopos pesados del xenón, cuyas masas varían entre 131 y 136, se enriquecían en hasta un 6 por ciento con respecto al xenón primordial.

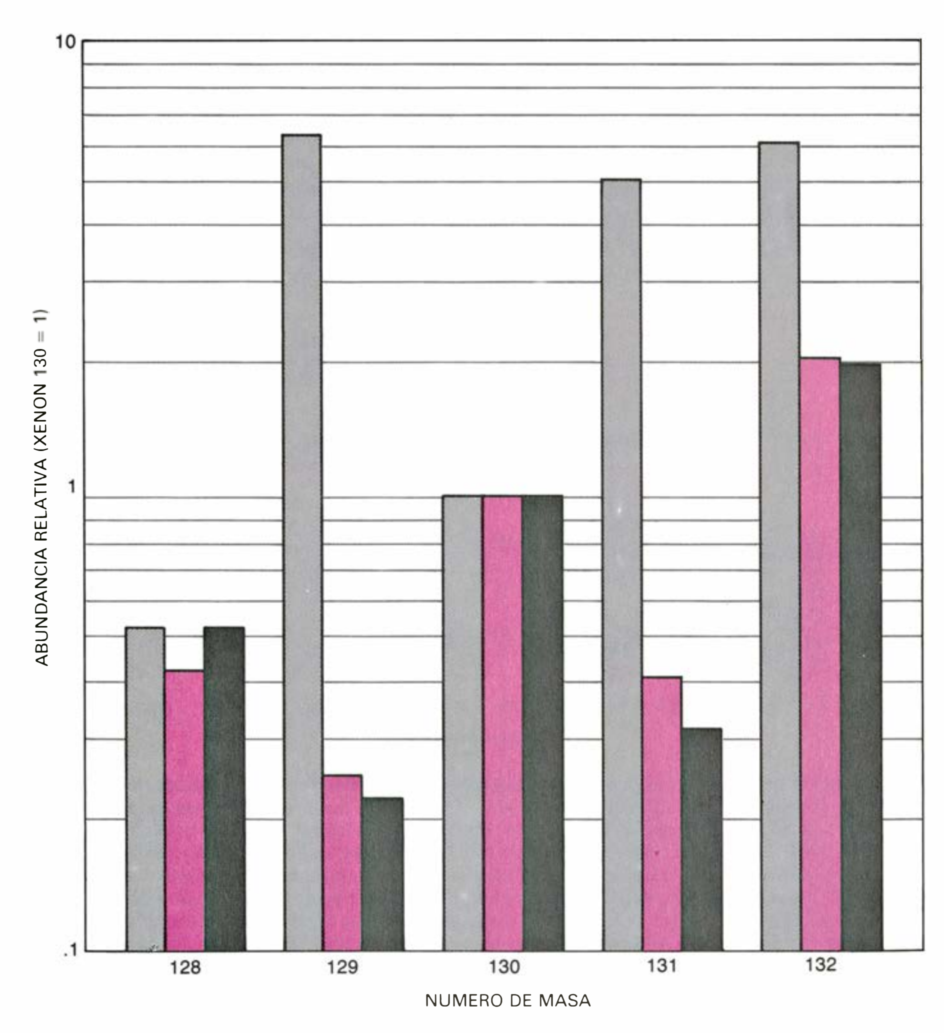
El enriquecimiento aumentaba del isótopo 131 al 136, como ocurre en el xenón formado por la fisión del uranio y otros elementos pesados. Reynolds y Turner sugirieron, en consecuencia, que el nuevo componente xenón provenía de la fisión de algún elemento pesado extinto que se dio, alguna vez, en el meteorito. También señalaron un problema: los raros isótopos ligeros xenón 124 y 126 también estaban enriquecidos en el meteorito. Ahora bien, estos isótopos no se forman por fisión, luego su enriquecimiento "debe explicarse mediante algún otro proceso". Al parecer, no uno, sino dos nuevos componentes de xenón están presentes en el meteorito. Se designaron después *H* y *L* (del inglés *H*eavy, pesado, y *L*ight, ligero). Aunque de distinto origen, han demostrado ser inseparables en los meteoritos, por cuya razón denominaremos su mezcla como xenón *HL*.

¿Un elemento superpesado?

Ninguno de los elementos fisionables conocidos más pesados que el uranio proporciona xenón con la configuración isotópica del xenón *H*. Además, la extensión de la tabla periódica a elementos más allá de 100 mostró que sus semividas se acortaban a medida que aumentaba el número atómico (es decir, el número de protones), lo que apuntaba que quizá no hubiera síntesis de nuevos elementos químicos más allá del 106 o 107. Pese a ello, W. D. Myers y W. J. Swiatecki, de Berkeley, trataron de calcular, en 1966, las propiedades de elementos todavía más pesados. Se encontraron con un resultado sorprendente: las semividas predichas se dilataban en vez de acortarse. De acuerdo con sus cálculos, los siguientes "números mágicos" (que corresponden a configuraciones particularmente estables de neutrones o protones en un núcleo) se producirían a 114 protones y



7. OTRO TIPO de xenón anómalo se descubrió en el calentamiento escalonado de un residuo del meteorito de Murchison. Srinivasan y uno de los autores (Anders) lo habían tratado con reactivos que oxidan los polímeros orgánicos. Se pretendía así eliminar el xenón primario. Las dos primeras fracciones de gas estaban todavía cercanas a la línea que representa mezclas del xenón primario y el xenón que Reynolds y Turner habían descubierto. Las fracciones siguientes se apartaban de esta línea. Las fracciones aberrantes demostraron contener una clase de xenón, *s*, compuesta principalmente de isótopos intermedios.



8. COMPOSICION ISOTOPICA del xenón *s* (*barras en color*). Aunque difiere de la del xenón solar (*barras gris claro*), se ajusta a la debida al proceso *s* (captura de neutrones lentos) en las estrellas gigantes rojas (*barras gris oscuro*). El cuadro del proceso *s* fue calculado por D. D. Clayton y R. A. Ward, de Rice.

184 neutrones, dando lugar a una "isla de estabilidad" en la tabla periódica centrada alrededor del elemento 114. Los intentos por sintetizar tales elementos superpesados en los aceleradores o identificarlos en los rayos cósmicos continúan. Sus semividas predichas, aunque altamente inciertas, varían hasta 10^8 años. En esa línea, tres grupos de investigadores de los Estados Unidos y Polonia propusieron, en 1969, que el elemento padre del xenón *H* podría ser un elemento superpesado extinto de la isla de estabilidad.

La conjetura se sometió en seguida a investigación. O. K. Manuel y sus colegas de la Universidad de Missouri examinaron el xenón de tres condritas carbonáceas. En cada una hallaron la misma intrigante correlación de isótopos ligeros y pesados que Reynolds y Turner habían observado, lo que constituía un serio obstáculo para la hipótesis de la fisión. Si el xenón *H* se produjo por fisión de un elemento pesado en el interior del meteorito, ¿por qué iba siempre acompañado de la misma proporción de xenón *L*, cuyo origen debía ser diferente?

Manuel y sus colegas llegaron a la conclusión de que ambos tipos de xenón se produjeron juntos antes de que se formasen los meteoritos y quedaron aprisionados en ellos. Sugirieron tres posibles orígenes, uno de los cuales era la nucleosíntesis en diferentes zonas de una supernova. El xenón *H* se produciría mediante el proceso *r* (captura de neutrones por núcleos en una rápida escala de tiempo) y el xenón *L* mediante el proceso *p* (captura de protones o algún otro proceso que diera lugar a isótopos pobres en neutrones).

Tropezando con el portador

Para someter a prueba las hipótesis (fisión dentro del portador por un lado, nucleosíntesis por otro) era necesario aislar el portador y estudiarlo con detalle. Si el xenón *H* fuera exótico, su portador debería también serlo y mostrar anomalías isotópicas propias. Pero si se produjo a través de un proceso de fisión en el meteorito, su portador mostraría señales de la misma: otros productos de fisión, por ejemplo. En colaboración con B. Srinivasan, ahora en la Universidad del estado de Washington, emprendimos el proyecto.

Como ocurre frecuentemente en nuestro trabajo, realizamos el experimento correcto con una idea errónea. Partiendo de la falsa sospecha de que el mineral pentlandita [$(Fe,NI)_9S_8$] contenía xenón *H*, disolvimos una muestra de la condrita carbonácea de Allende en ácido clorhídrico y ácido fluorhídrico, que se supone no disuelven la pentlandita. Quedó un residuo negro que representaba aproximadamente 0,5 por ciento de la masa del meteorito y que contenía la mayor parte del xenón *H*. Pero la pentlandita resultó ser sólo una pequeña parte del residuo y, además, carecía de importancia. A fin de cuentas, el residuo demostró estar constituido por carbono amorfo, cromita (Cr_2FeO_4) y espinela.

Y algo que provocó nuestra irritación: el residuo contenía abundante xenón planetario, que enmascaraba por completo el *H*. Para fortuna nuestra cometimos entonces un segundo error. Todavía pensando que al menos uno de los componentes de xenón estaba aprisionado en un sulfuro, tratamos el residuo con ácido nítrico, que disuelve todos los sulfuros. A buen seguro, se perdió el xenón primitivo en el residuo mientras que quedó el xenón *HL* en forma mucho más pura que en ningún caso anterior. La verdad es que en un sulfuro no queda aprisionado nada de xenón. El experimento tuvo éxito por una razón distinta. El xenón primordial queda aprisionado en la superficie de los granos, y se pierde cuando el ácido nítrico ataca la superficie. El xenón *HL* queda aprisionado dentro de los granos y allí permanece.

En este punto, pues, nos las habíamos arreglado para concentrar el portador o portadores de xenón *HL* en un residuo que representaba una pequeña fracción de la masa del meteorito. Podíamos buscar ya anomalías isotópicas debidas a la nucleosíntesis estelar o a la fisión. Las primeras pruebas resultaron negativas: tanto el carbono como el osmio eran normales. Robert Clayton y Mark H. Thiemens en Chicago y Urs Frick y Pepin en Minnesota probaron entonces con el oxígeno y el nitrógeno. Su carácter anómalo era manifiesto. Mas como el residuo contenía todavía una mezcla de carbono, cromita y espinela, no quedaba claro si las anomalías estaban asociadas con los portadores reales del xenón. La mezcla de minerales recordaba el juego de muñecas rusas metidas una dentro de otra. Si el xenón *HL* era la primera muñeca, la más interna, de nada serviría medir la composición isotópica de otros elementos en las muñecas subsiguientes. ¿Cómo hallar la muñeca más interna y tomar en ella todas las medidas?

En la fase siguiente se eliminó la espinela, habida cuenta de que la espinela pura extraída del residuo no tenía prácticamente gases. Quedaban el carbono y la cromita. Pero un estudio detallado de esos minerales, hecho en parte por Frick y Ulrich Ott, de Berkeley, y Sherwood Chang, del Centro de Investigación Ames de la Administración Nacional de Aeronáutica y del Espacio, mostraron que la mayor parte del carbono y la cromita eran pobres en gases aprisionados. Casi todo el xenón *HL* estaba, pues, en una pequeña parte del carbono (y posiblemente cromita),

9. ENTORNO DEL XENON en una tabla de isótopos, donde se muestra cómo los isótopos se producen a través de procesos distintos: el proceso *s* (*s*), el *r* (*r*) y el *p* (*p*). Cada uno lleva consigo movimientos característicos en este tablero de ajedrez nuclear. En el proceso *s*, los núcleos atómicos estables capturan neutrones lentamente (*líneas negras horizontales*). Cuando una captura hace un núcleo inestable (esto es, radiactivo) (*casillas blancas*), tiene lugar una desintegración beta (*línea negra diagonal*): el núcleo emite un electrón, convirtiendo uno de sus neutrones en un protón. El proceso *s* no puede producir los dos isótopos más ligeros ni los dos más pesados del xenón; con ello se ajusta a la configuración del xenón descubierto por Srinivasan y Anders. En el proceso *r*, o captura de neutrones rápidos, los núcleos ricos en neutrones (la mayoría fuera de la escala del diagrama, hacia la parte inferior derecha) se producen con tanta rapidez que los propios radiactivos de vida corta tienen poca oportunidad de desintegrarse. Cuando se ha agotado la provisión de neutrones, una sucesión de desintegraciones beta los convierte en isótopos estables. Se muestra una serie, que conduce al xenón 136 (*línea de flechas diagonales en color*). Los núcleos radiactivos producidos por fisión se desintegran según esquemas similares. Algunos isótopos estables (*casillas en color claro*) no pueden producirse mediante desintegración beta. Están "defendidos": un núcleo estable más rico en neutrones (*casillas en color vivo*) termina la sucesión. Los isótopos defendidos se forman en el proceso *s* o el *p*: captura de protones o expulsión de neutrones por rayos gamma.

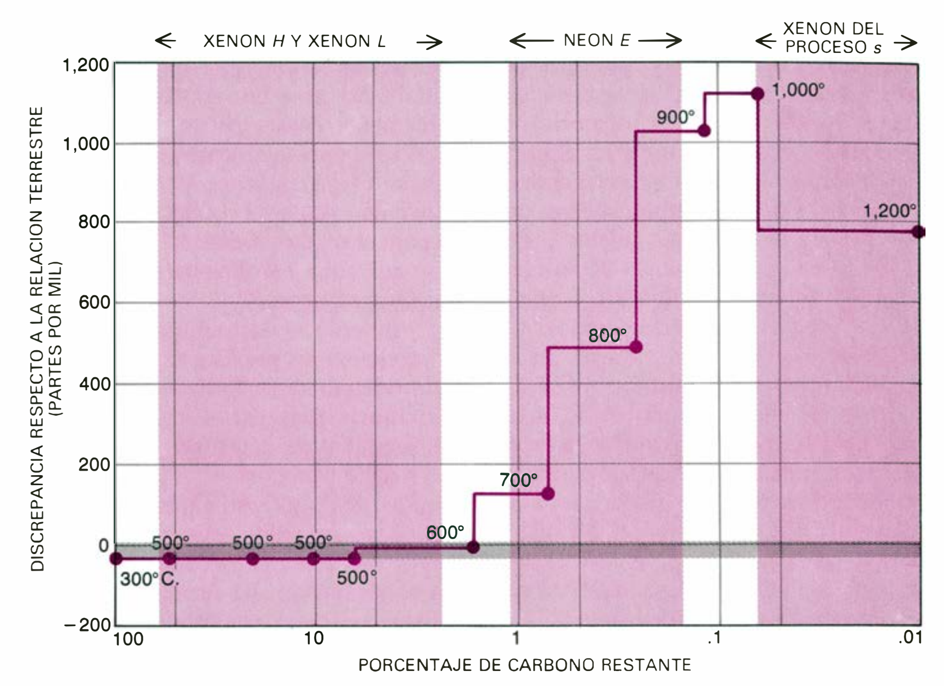
que podía ser o no la muñeca rusa más interna.

El año pasado, P. K. Swart, M. M. Grady y C. T. Pillinger, de la Universidad de Cambridge, determinaron la composición isotópica del carbono. Hallaron que la relación de carbono 12 a carbono 13 era 93, número que está dentro del margen del carbono terrestre (88 a 93). A primera vista, este resultado apoya el origen local del carbono y consiguientemente del xenón, ya que una relación de 93 es muy poco corriente en la galaxia. Los valores en las estrellas varían entre menos de 10 y varios centenares; el promedio para las nubes interestelares es de 62.

Dificultades para la fisión

Los ensayos destinados a averiguar si hubo fisión no aportaron mucha luz. Y así la situación de estancamiento persistió durante más de 10 años, hasta que se realizaron dos experimentos cruciales en 1982. En el primero de ellos, nosotros, trabajando junto con I. P. Wright, S. J. Norris y Pillinger, de Cambridge, medimos la composición isotópica del nitrógeno en dos residuos de la condrita carbonácea de Murchison, uno de los cuales, el de grano más fino, estaba enriquecido casi 600 veces en neón *HL*. El residuo demostró contener un nitrógeno muy poco corriente, enriquecido en nitrógeno 14 casi en un 30 por ciento. Tal anomalía sólo podía resultar de una nucleosíntesis estelar y no de fisión, de modo que si la muestra de grano fino era la muñeca rusa más interna, el xenón *HL* debía haberse producido también en una estrella. Nos quedábamos sin saber por qué el carbono de la muestra era tan perfectamente normal. Uno se maravilla del sentido del humor que manifiesta la naturaleza al escoger carbono de composición isotópica terrestre para envolver nitrógeno y xenón altamente anómalos.

En el segundo experimento, los autores, trabajando con G. W. Lugmair y Tadashi Shimamura, de la Universidad de California en San Diego, sometieron a prueba el meteorito de Allende en busca de anomalías isotópicas en el bario, neodimio y samario. Cualquier proceso que fabricase xenón *HL* debería haber producido cantidades comparables de al menos algunos isótopos de esos elementos vecinos en proporciones que variarían de un proceso a otro. No encontramos anomalías de ninguna clase (excepto enriquecimientos de neodimio 142 y 143, ambos atribuibles a desintegración radiactiva). Este descubrimiento parece, a primera vista, cuestionar por igual las dos hipótesis



10. COMBUSTION DE UN RESIDUO DE CARBONO del meteorito de Murchison. Así deja ver las tres clases exóticas de carbono que esconde. Cada una transporta neón o xenón anómalos. La muestra se calentó en escalones en presencia de oxígeno; en cada fracción resultante de dióxido de carbono se determinó la relación de carbono 13 a carbono 12. Las primeras fracciones procedieron del portador de xenón *H* y xenón *L*. Por coincidencia, su carbono tenía proporciones isotópicas como las del carbono terrestre (*banda gris*). Las fracciones siguientes procedieron de los portadores de neón *E* y xenón *s*. Dicho experimento fue realizado por P. K. Swart, M. M. Grady y C. T. Pillinger, de la Universidad de Cambridge.

acerca del origen del xenón *H*, es decir, la nucleosíntesis estelar y la fisión en el interior del meteorito. Para la nucleosíntesis, la química ofrece una posible salida a la dificultad: el bario, el neodimio y el samario no son nada volátiles y deberían hallarse, por tanto, entre los primeros elementos que se condensasen en el gas en expansión y en enfriamiento expelido por una supernova. Presumiblemente, se condensarían sobre óxidos, tales como la perovskita ($CaTiO_3$). El xenón, gas noble, se condensaría mucho más tarde, a temperaturas más bajas y sobre un substrato diferente, carbono quizá.

Al no haber una excusa similar para la hipótesis de la fisión, nos quedamos con la nucleosíntesis estelar. Cabe la posibilidad de alguna variante del proceso *r* durante la "combustión" explosiva del carbono, que puede producirse en una supernova cuando una onda de choque atraviesa una zona rica en carbono de la estrella moribunda. Los cálculos hechos por Dieter Heymann y Marlene Dziczkaniec, de Rice, sugieren que los neutrones y los rayos gamma (fotones de alta energía), liberados a temperaturas de 2000 millones de grados, vuelven a elaborar la composición isotópica de los elementos pesados en la zona; el xenón emergería así en la proporción isotópica correcta. Y cabe otra posibilidad: un proceso *r* más rápido que forma núcleos de todas las masas hasta los elementos superpesados de corta vida de masa cercana a 280, los que después se fisionan en elementos más ligeros, entre ellos el xenón. Los cálculos realizados por E. P. Steinberg y B. D. Wilkins, del Laboratorio Nacional de Argonne, muestran que la fisión ayudaría a explicar la gran abundancia, en el sistema solar, de los elementos de tierras raras terbio, disprosio, holmio y erbio. También justifican de una manera plausible los datos del xenón *H*.

Si existen o no los elementos superpesados de vida corta se decidirá, a buen seguro, cuando concluyan los trabajos empeñados en su síntesis que se llevan a cabo en los centros más importantes de investigación nuclear en Berkeley, Dubna en la Unión Soviética y Darmstadt en Alemania. Aunque la perspectiva favorable a la existencia de elementos superpesados de vida larga se ha oscurecido bastante a raíz de los últimos estudios sobre el xenón *H*.

¿Granos de polvo de las estrellas gigantes?

Mientras luchábamos con el xenón *HL*, descubrimos accidentalmente un tercer tipo de xenón anómalo. Más raro y mejor escondido que los demás, su composición isotópica denuncia con nitidez su lugar de origen.

De nuevo hicimos el experimento co-

rrecto a pesar de la idea errónea que nos movía. Srinivasan y uno de nosotros (Anders) pretendíamos caracterizar el xenón *HL* en el meteorito de Murchison confiados en que diferiría del xenón de Allende en su cociente de xenón *L* a xenón *H*. (Murchison es una condrita carbonácea C2, mientras que Allende es una C3.) Nos contrarió que los residuos atacados de Murchison poseyeran todavía una gran cantidad de xenón primordial. Presumiblemente, el tratamiento con ácido nítrico había sido incapaz de eliminar todo el polímero orgánico. (El polímero orgánico contiene gases primodiales y está presente en las condritas C2 pero no en las C3.) Tratamos los residuos con NaOCl y Na_2O_2, que oxidan los polímeros orgánicos. Y examinamos las muestras mediante el calentamiento escalonado.

Los dos primeros datos revelaron lo que estábamos buscando. El xenón primordial se desprendió a 800 grados Celsius, seguido a 1000 grados por el xenón *HL*, más puro que el que residuo Murchison alguno había proporcionado hasta entonces. Los cuatro puntos siguientes se conviertieron en sendas sorpresas. Empezando a 1100 grados, se desviaron de la línea que en nuestro diagrama representaba mezclas de xenón primordial y xenón *HL*. Al parecer, la muestra contenía un nuevo componente de xenón que se encontraba en el punto 1600 grados, o más allá de él. Constituye menos de una parte en 10^4 de xenón total; se nos había escapado en todos los ensayos anteriores.

El nuevo componente se mostró enriquecido, sobre todo, en los isótopos pares del xenón: el 128, 130 y 132; presentaba, por contra, fuerte déficit de los isótopos del xenón más ligeros y más pesados: 124, 126, 134 y 136. Nos hallábamos ante una configuración notablemente característica, que denunciaba un evidente proceso *s*. Pura coincidencia, Donald Clayton y R. Ward acababan de calcular la configuración teórica para el xenón procedente de procesos *s*: concordaba cabalmente con nuestro nuevo componente.

El proceso *s*, como el proceso *r*, fabrica elementos pesados por captura de neutrones, aunque lo hace de una manera mucho más lenta: entre dos capturas sucesivas de neutrones por un mismo núcleo transcurren años, no milisegundos. De aquí que cualquier núcleo radiactivo de vida corta que se forme tenderá a desintegrarse antes de capturar otro neutrón. La formación de núcleos sigue, pues, un "camino de proceso *s*" que lleva consigo capturas de neutrones intercalados con desintegraciones. Tomemos el telurio 122, un isótopo estable. El proceso *s* lo convierte secuencialmente en telurio 123, 124, 125 y 126, todos los cuales son estables o no radiactivos. Viene a continuación el telurio 127, isótopo radiactivo, que se desintegra dando iodo 127. (Específicamente, un neutrón en el núcleo del telurio 127 emite una partícula beta, o electrón de alta energía y se convierte por ello en un protón.) La captura de otro neutrón produce iodo 128, radiactivo; después pasa a xenón 128, a través del isótopo 132; a xenón 133, que es radiactivo; a cesio 133, y así sucesivamente. Nótese que el proceso *s* no produce los isótopos del xenón más ligero y más pesado, que son precisamente los que no se encuentran en el nuevo componente del meteorito.

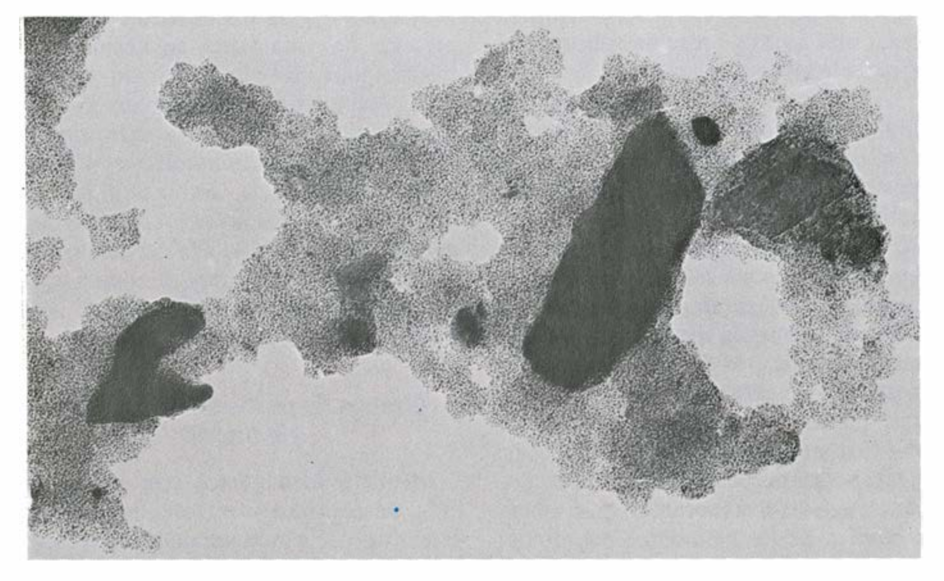
El proceso *s* se desarrolla en las gigantes rojas: estrellas que han agotado el hidrógeno en su núcleo y lucen mediante reacciones termonucleares que convierten helio 4 en carbono 12. Consumen también el hidrógeno que queda en una capa gaseosa más hacia el exterior. El proceso *s* se desarrolla bajo la capa de combustión del hidrógeno, donde los neutrones se liberan mediante reacciones del estilo de la que funde el neón 22 y el helio 4 para dar magnesio 25, permitiendo el escape de un neutrón. El hierro 56 y otros núcleos pesados capturan entonces los neutrones; así comienza el proceso *s*.

El xenón originado en el proceso *s* (xenón *s* para abreviar) debió pasar algún momento de su historia en el interior de una estrella gigante roja. ¿Cuándo y cómo escapó? Quizás en la fase de gigante roja. Se sabe que las estrellas gigantes rojas emiten un viento de partículas a través del cual pierden masa a un ritmo de 10^{-7} veces la masa del Sol por año. Tal vez en algún momento posterior, cuando una estrella gigante roja evoluciona hasta convertirse en nova o supernova. La materia es expulsada en todas las fases de la evolución de una estrella gigante roja y se enfría por expansión y radiación hasta que su temperatura es suficientemente baja (menos de 200 grados Kelvin) para que se condensen los granos. Los isótopos estables del xenón y otros elementos pesados, que en otra época se habían formado mediante el proceso *s*, no sufren alteración ulterior en una gigante roja hasta que la estrella se convierte en una supernova y las temperaturas aumentan por encima de los 2000 millones de grados Kelvin. En contraste, los elementos más ligeros tales como el carbono, nitrógeno, litio, helio y neón continúan reaccionando a medida que la estrella evoluciona y se calienta. Nos centramos en los elementos más ligeros. Su composición isotópica podía proporcionar alguna pista sobre la fase exacta de expulsión de los granos que transportaban xenón *s*.

Carbono pesado

No tardamos en encontrar que el xenón *s* de los meteoritos estaba en el carbono. Tratamos nuestro residuo del meteorito de Murchison con ácido perclórico, que oxida el carbono a dióxido de carbono y deja inalterados la mayoría de los minerales. El xenón *s* desapareció.

Había motivos para esperar que este carbono fuese isotópicamente anómalo, porque las gigantes rojas y sus estrellas sucesoras tienen relaciones de car-



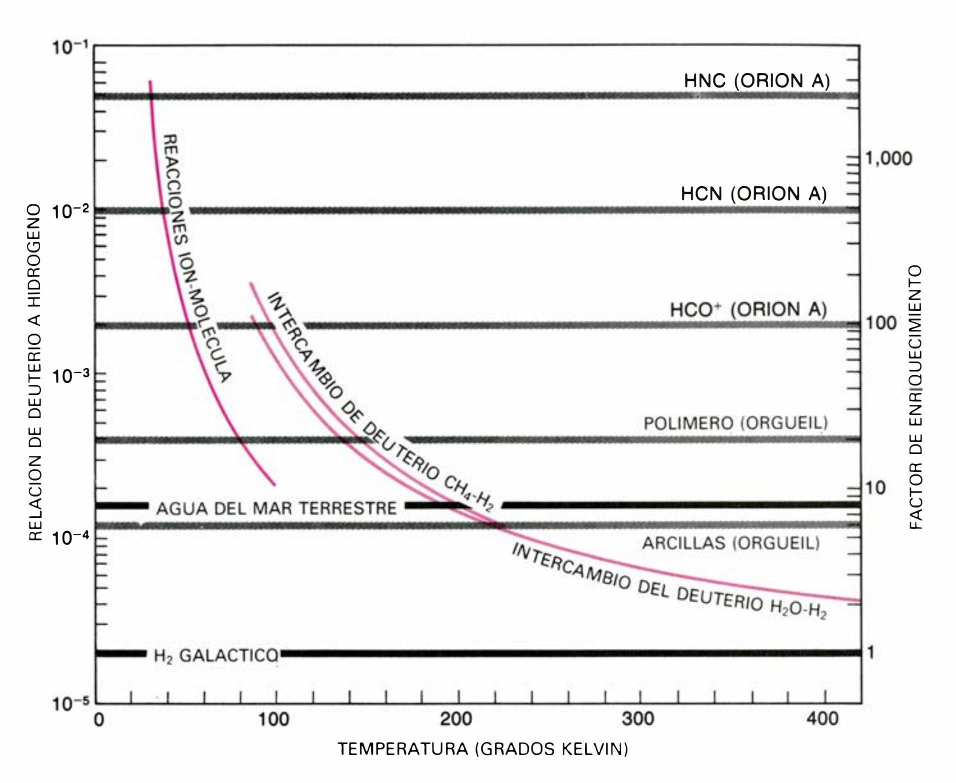
11. MICROGRAFIA ELECTRONICA de una muestra de Murchison que contiene tres tipos exóticos de carbono. La tomó Mitsuo Ohtsuki, de la Universidad de Chicago. Solamente se puede reconocer una de las clases: la materia de grano fino transporta xenón *H* y xenón *L*. Las otras dos clases constituyen un pequeño porcentaje de los granos mayores y más oscuros que alcanzan tamaños de hasta un micrometro.

bono 12 a carbono 13 que varían en más de un centenar. El problema era cómo aislar este carbono exótico de la cantidad mucho mayor de carbono ordinario que presumiblemente se encontraba todavía en la muestra. Usamos el residuo de grano fino de Murchison que se había analizado con respecto al nitrógeno. Se le había desmineralizado con ácidos, atacado con agentes oxidantes alcalinos para eliminar el polímero y, finalmente, ultrafiltrado para retener materia de diámetro superior a un micrometro. De su análisis se encargó Swart, Grady y Pillinger.

Para separar los componentes de carbono, siguieron la técnica de combustión escalonada inventada por David J. Des Marais, del Centro de Investigación Ames: a temperaturas de progresión creciente se calienta una muestra con oxígeno, y se analiza en un espectrómetro de masas el dióxido de carbono formado en cada escalón. Las diferentes clases de carbono de la muestra se queman según su tamaño de grano y cristalinidad; con un poco de suerte se analizan una a una.

La mayor parte del carbono pertenecía al portador del xenón *HL*, cuya composición característica tenía: alrededor de un 3 o 4 por ciento más ligero que el carbono ordinario. Sin embargo, por encima de 600 grados Celsius el carbono se volvió progresivamente más pesado, alcanzando un 110 por ciento de enriquecimiento en carbono 13 entre 900 y 1000 grados. No había precedentes de una composición tan extrema. Al parecer, el meteorito encierra un nuevo tipo de carbono isotópicamente más pesado. Si el enriquecimiento más extremo de 110 por ciento representa este carbono pesado en forma pura, nuestro residuo de Murchison contendrá un 0,45 por ciento de carbono pesado; el grueso del meteorito posee unas cinco partes por millón.

El carbono pesado es, probablemente, mezcla de dos tipos de carbono: el portador del xenón *s* y el del neón *E*(*L*). El portador del xenón *HL* es también carbono, lo que da un total de tres clases de carbono exótico y presumiblemente presolar en los meteoritos primitivos.



12. EL DEUTERIO ANOMALO revela una cuarta clase exótica de carbono. El diagrama compara la relación de deuterio a hidrógeno ligero (esto es, de hidrógeno 2 a hidrógeno 1) para cierto número de fuentes de hidrógeno. Se toma, por término de comparación, la relación en nuestra galaxia. El agua de mar terrestre está enriquecida en la proporción de ocho veces. Las arcillas del meteorito de Orgueil lo están en proporción ligeramente menor; el polímero de carbono de Orgueil aparece más enriquecido. Las moléculas de las nubes interestelares, Orión A por ejemplo, están enriquecidas hasta 100.000 veces. En principio, las reacciones que transfieren deuterio de moléculas de hidrógeno a otras moléculas (*color claro*) podrían establecer el enriquecimiento; las reacciones, sin embargo, son lentas. Más probable es que el polímero lleve consigo indicios de moléculas interestelares, cuyo enriquecimiento se produce cuando iones moleculares tales como el CH_2^+ reaccionan con moléculas de hidrógeno (*color oscuro*). La comparación de los datos fue ideada por J. Geiss, de la Universidad de Berna, y Hubert Reeves, del CENS.

Trazador de moléculas interestelares

La última pista que seguiremos atañe al deuterio (hidrógeno 2). Conduce a una cuarta clase de carbono exótico. El primer indicio apareció en 1953. Giovanni Boato, quien trabajaba en el laboratorio de Harold C. Urey en Chicago, halló varias condritas carbonáceas enriquecidas en deuterio hasta un 31 por ciento (con respecto al agua del océano). Transcurrieron años sin que se prestara mayor atención a la pista del enriquecimiento, responsabilizando de él a una de las principales formas de hidrógeno en las condritas carbonáceas: los minerales de arcilla que contienen grupos OH y materia orgánica que contiene grupos CH (principalmente el polímero orgánico mencionado antes).

El experimento clave llegaría en 1979. Y. Kolodny, J. F. Kerridge e I. R. Kaplan, de la Universidad de California en Los Angeles, estimulados por el trabajo de F. Robert y sus colegas en el Centro de Saclay de Investigación Nuclear (CENS), analizaron muestras de condritas carbonáceas por calentamiento escalonado antes y después de que la materia orgánica en las muestras se hubiera quemado convirtiéndose en un plasma (gas ionizado) de oxígeno. A partir de la diferencia entre los dos análisis dedujeron el contenido de deuterio de la materia orgánica. El enriquecimiento en deuterio se concentraba en la materia orgánica, con excesos que llegaban hasta el 160 por ciento. Medidas posteriores de Robert, Richard H. Becker, Jiyoung K. Yang y Samuel Epstein, del Instituto de Tecnología de California, extendieron el dominio hasta el 310 por ciento; pusieron de manifiesto que, de las varias clases de materia orgánica, el polímero evidenciaba mayor enriquecimiento.

De todos los isótopos, el deuterio es el más propenso al fraccionamiento, pues su masa duplica la de su isótopo hermano, el hidrógeno 1. De cualquier modo, no es fácil justificar un enriquecimiento del 310 por ciento, pues representa un factor de 4,1. La verdad es que el enriquecimiento supera incluso ese valor, porque el agua del mar en sí ya está enriquecida en ocho veces respecto al hidrógeno galáctico. Así, el enriquecimiento global del polímero meteorítico es de 32 veces.

En principio, el enriquecimiento podría producirse por reacciones de intercambio isotópico, tal como la que transfiere un átomo de deuterio (D) del hidrógeno molecular al metano (CH_4): $CH_4 + HD = CH_3C + H_2$. Según los cálculos teóricos, esta reacción enriquece el contenido de deuterio del metano en 32 veces, a una temperatura de unos 130 grados Kelvin. Enriquecimientos similares se producen cuando

CLASE DE CARBONO	CLASE DE METEORITOS	CONCENTRACION DE CARBONO (PARTES POR MILLON)	GASES NOBLES U OTROS TRAZADORES	ORIGEN	RELACION DE CARBONO 12 A CARBONO 13
ORGANICO	C1, C2	APROXIMADAMENTE 30.000	PRIMORDIAL	LOCAL ($CO + H_2$)	91
ELEMENTAL	C3	2000	PRIMORDIAL	LOCAL ($2CO \rightarrow C + CO_2$)	91
CARBONATOS	C1, ALGO EN C2	2000 O MENOS		LOCAL ($CO_2 + MgO$)	83-85
CARBONO ALFA	C1, C2 PARTE EN C3	5 O MENOS	NEON *E*	¿NOVA?	80 O MENOS
CARBONO BETA	C1, C2 PARTE EN C3	5 O MENOS	XENON DEL PROCESO *s*	GIGANTE ROJA	42 O MENOS
CARBONO DELTA	C1, C2, C3	200	XENON *H* XENON *L* Y NITROGENO LIGERO	SUPERNOVA	92
ORGANICO	C1, C2	3000 O MENOS	DEUTERIO	NUBE MOLECULAR INTERESTELAR	?

13. CLASES DE CARBONO en las condritas carbonáceas. Tres tipos son locales; se explican por procesos ocurridos en el primitivo sistema solar. Los cuatro restantes son exóticos: su origen interestelar queda revelado por la composición isotópica de su carbono o por anomalías isotópicas en "trazadores".

se forma materia orgánica a partir de pequeñas moléculas. La dificultad estriba en que las velocidades de reacción a 130 grados son ridículamente pequeñas. El tiempo medio para la formación del metano a partir del monóxido de carbono e hidrógeno molecular es de 10^{30} años.

J. Geiss, de Berna, y Hubert Reeves, de Saclay, han avanzado una explicación mejor, fundados en una propuesta que adelantaron en 1972. Las moléculas del espacio interestelar, HCHO y HCN por ejemplo, están enriquecidas en deuterio en factores de hasta 10^5, según sus radioespectros. Tal vez los meteoritos contienen indicios de tales moléculas interestelares.

Reacciones ion-molécula

¿Por qué están las moléculas interestelares enriquecidas en deuterio? La pregunta la respondería, en 1973, William D. Watson, de la Universidad de Illinois en Urbana-Champaign. Se sabía ya que las moléculas interestelares se formaban principalmente por una cadena de reacciones ion-molécula, desarrolladas con rapidez incluso a las temperaturas más bajas. Un ion carbono producido por los rayos cósmicos reacciona con una molécula de hidrógeno para formar el ion molecular CH_2^+. El ion molecular reacciona entonces con otras moléculas de hidrógeno para constituir un ion más complejo. Por último, reacciona con un electrón libre para formar una molécula estable y neutra, el CH_4. Watson observó que el enlace químico entre el carbono y el deuterio era ligeramente más fuerte que el enlace entre carbono e hidrógeno 1, lo que determinaba que el deuterio se concentrase en iones durante las reacciones de intercambio. Por la misma razón, los iones que desprenden hidrógeno tienden a perder hidrógeno 1, en vez de deuterio, lo que provoca que el deuterio se concentre todavía más en los iones moleculares y, finalmente, en la molécula neutra. El grado de enriquecimiento depende de la molécula particular, pero aumenta con temperatura decreciente.

En los meteoritos, la materia rica en deuterio aparece en forma de un polímero orgánico complejo, y no como pequeñas moléculas primitivas del tipo de las detectadas por los radioastrónomos. Ello no debe sorprendernos, ya que tales moléculas son demasiado volátiles y demasiado reactivas para persistir sin cambios. El carbono del polímero no es isotópicamente excepcional (1,6 por ciento más ligero que el carbono normal), lo que sugiere que el carbono de las moléculas interestelares no es excepcional, o, lo que es más probable, que el carbono interestelar está grandemente diluido por el carbono "local" o primario.

El polímero local tiene, en cualquier caso, una estructura complicada: sistemas de anillos aromáticos con hasta cuatro de ellos fundidos y soldados mediante puentes de hidrocarburo $[(CH_2)_n]$. Esto no nos aclara su origen; se asemeja al carbono formado de materia vegetal. Al parecer, muchas clases de materia orgánica constituyen polímeros aromáticos bajo calor persistente. La parte interestelar del polímero, rica en deuterio, no necesita tener la misma estructura, porque muchas moléculas interestelares, incluyendo el HNC, $H(C \equiv C)_nCN$ y HCHO, son muy reactivas y pueden haberse polimerizado en la superficie de granos a bajas temperaturas. Tal vez esta diferencia permitirá separar los polímeros locales y exóticos por procedimientos químicos.

Carbono del sistema solar

En el curso de este artículo se han examinado cuatro clases de carbono exótico de los meteoritos primitivos. Tres son formas de carbono elemental portador de neón $E(L)$, xenón HL y xenón s; el cuarto, un polímero enriquecido en deuterio. Cada uno de ellos está escondido en el carbono local. ¿Qué llevó al carbono a tal riqueza de estados químicos?

Podemos tratar de responder la pregunta en términos de termodinámica, que predice los compuestos de carbono que deberían haberse formado en la nebulosa solar a diversas temperaturas y presiones. Veamos pues lo que ocurre cuando un gas solar se enfría desde altas temperaturas a una presión de 10^{-5} atmósferas, valor razonable para la región del futuro cinturón de asteroides. (El modelo que presentaremos está basado en trabajo realizado por Ryoichi Hayatsu y uno de los autores, Anders, con M. H. Studier, de Argonne.) En un comienzo, el carbono se presenta principalmente en forma de monóxido de carbono gaseoso, tipo dominante del carbono en el espacio interestelar. Queda en esa forma mientras el gas se contrae y se calienta hasta constituir la nebulosa solar. Al enfriarse, debería hidrogenarse produciendo metano. Sin embargo, en ausencia de catalizadores, la velocidad de reacción es muy baja. Lo mismo se aplica a otra reacción viable a temperaturas algo más bajas y convierte el monóxido de carbono en dióxido de carbono y carbono elemental.

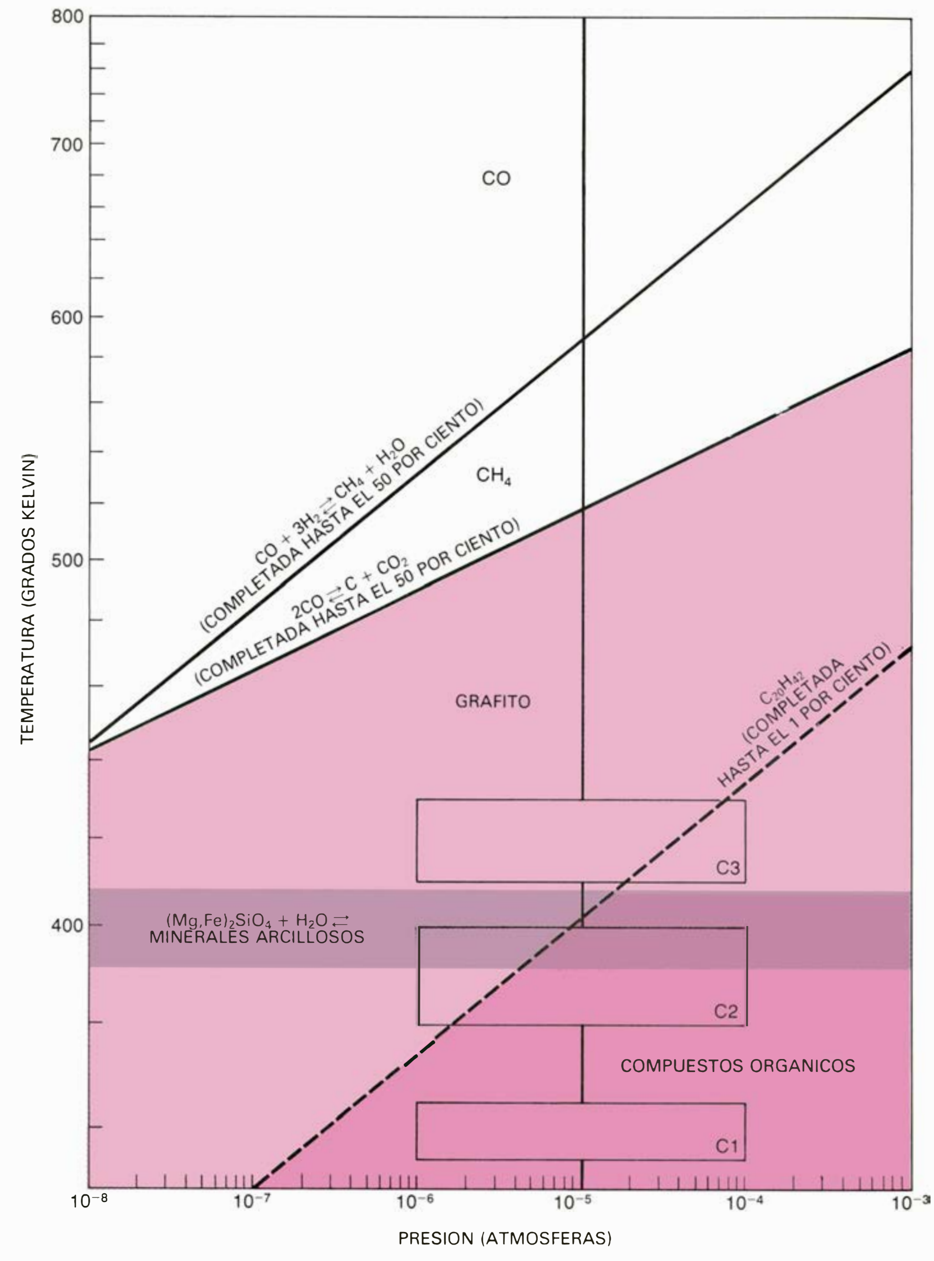
Por debajo de 400 grados Kelvin, la situación cambia de raíz. Se forman minerales de arcilla y magnetita, excelentes catalizadores ambos para la hidrogenación del monóxido de carbono; ahora bien, a esas temperaturas más bajas, el metano, el producto más estable de la hidrogenación todavía, no es ya el único posible. Además de otros muchos compuestos orgánicos, pueden surgir hidrocarburos superiores, por ejemplo $C_{20}H_{42}$. Pueden y lo hacen; el desarrollo de la reacción favorece el crecimiento de largas cadenas de hidro-

carburos. La reacción es esencialmente el proceso industrial Fischer-Tropsch: conversión de monóxido de carbono e hidrógeno en hidrocarburos, alcoholes y otros compuestos en presencia de un catalizador. El proceso industrial se desarrolló en 1923. Los hidrocarburos meteoríticos muestran el sello de la reacción: predominio de cadenas carbonadas lineales.

La secuencia de la nebulosa solar explica la diversidad de carbono local hallado en las condritas carbonáceas. Las condritas C3, que según los cálculos de J. W. Larimer, de la Universidad del estado de Arizona, se formaron a temperaturas entre 410 y 430 grados Kelvin, deberían contener mucho carbono amorfo, pero no compuestos orgánicos; y este es el caso. Las condritas C1 y C2, nacidas a temperaturas de menos de 400 grados, deberían contener principalmente carbono orgánico, y así sucede. Las dos formas de carbono orgánico hallado en las condritas C1 y C2 (compuestos solubles y polímero insoluble en los disolventes orgánicos corrientes) pueden representar diferentes duraciones de contacto entre la materia orgánica y el catalizador mineral. Las reacciones Fischer-Tropsch sólo dan, al principio, compuestos solubles en el laboratorio; transcurridos seis meses, estos compuestos se transforman en materia insoluble compleja, que se parece al polímero meteorítico. La pequeña cantidad de carbonato de los meteoritos pudo formarse cuando los óxidos de magnesio, calcio y hierro (productos secundarios de la formación de minerales de arcilla) reaccionaron con dióxido de carbono (un producto secundario de la reacción Fischer-Tropsch).

Para el carbono presolar hay una posibilidad más: la condensación directa de granos de grafito a alta temperatura. Los cálculos termodinámicos muestran que tal condensación sólo cabe en un gas cuya relación de carbono a oxígeno supere el valor 0,9, por encima de la relación solar de 0,6. Las gigantes rojas y sus estrellas sucesoras tienden a ofrecer tales relaciones. Además, hay indicaciones de que las capas de polvo que rodean a las estrellas gigantes contienen granos de grafito. Tales granos deben estar mejor cristalizados, y ser así más resistentes a cambios químicos, que la variedad de carbono amorfo que se produce a temperaturas más bajas a partir del monóxido de carbono. Ahí puede residir la razón de que purificáramos carbono mediante técnicas de combustión parcial y oxidación química. Sin embargo, en el proceso perdimos, probablemente, otros tipos reactivos de carbono.

Sin más herramientas que los cuantos de luz y las leyes de la física, los astrónomos y astrofísicos han deducido la naturaleza del funcionamiento interno de una abundante variedad de objetos astronómicos: gigantes rojas, novas, supernovas y nubes interestelares, entre ellos. Van surgiendo nuevas pistas: trozos de materia tangible de los objetos mismos, cada una de las cuales contienen un rico historial de la nucleosíntesis estelar y de la química interestelar. La cosecha potencial de conocimientos es grande, ya que las medidas de laboratorio pueden revelar muchos detalles que son inaccesibles a las técnicas astronómicas. Hace quince años no había esperanza siquiera remota de estudiar el polvo estelar; hoy, sus partículas están al alcance de la mano.



14. QUIMICA DEL CARBONO en la nebulosa solar. Nos explica los distintos estados del carbono local en las condritas carbonáceas. A altas temperaturas, el monóxido de carbono (CO) domina en la nebulosa. Más tarde, ésta se enfría. El CO debería convertirse en metano (CH_4); una línea inclinada muestra dónde la reacción está a medio completar. Sin embargo, en ausencia de catalizadores la reacción es lenta, de modo que la mayor parte del CO sobrevive. A continuación debería aparecer el dióxido de carbono (CO_2) junto con carbono elemental (C) en forma de grafito o granos amorfos de carbono. De nuevo la conversión es lenta. Cerca de los 400 grados Kelvin comienza un cambio drástico. Los silicatos reaccionan con agua, produciendo minerales arcillosos que catalizan la formación de compuestos orgánicos complejos. Una línea representa las combinaciones de temperatura y presión a las cuales la conversión de monóxido de carbono e hidrógeno en $C_{20}H_{42}$, típica molécula compleja, se ha completado en un 1 por ciento. Los tres rectángulos de la ilustración muestran los dominios de temperatura y presión en los que presumiblemente surgieron las condritas carbonáceas. Las condritas C3, tales como el meteorito de Allende, se formaron a temperaturas superiores a 400 grados Kelvin: contienen pequeñas cantidades de carbono, principalmente amorfo. Las condritas C1, así Orgueil, y C2, Murchison por ejemplo, nacieron a temperaturas más bajas. Contienen compuestos orgánicos y, también, los minerales arcillosos que actuaron de catalizadores.

Las defensas químicas de los termes

De cuerpo blando y ciegos, los termes se hallan expuestos a la depredación. En su defensa, los termes soldados atacan a los invasores con un arsenal de irritantes, anticoagulantes y viscos

Glenn D. Prestwich

Desde hace más de 100 millones de años, el mundo de los insectos se halla implicado en una carrera de armamentos químicos. Los depredadores intentan someter a sus víctimas con venenos tóxicos o atraerlas con perfumes seductores. Las presas responden mediante la eyección de sustancias irritantes, pegajosas, calientes o venenosas, o bien convirtiéndose en incomestibles. Entre los insectos con defensa química sobresalen los soldados de los termes, que constituyen una casta dentro de cada especie y cuya cabeza y cuerpo se hallan de tal forma modificados en armas que no pueden alimentarse por sí mismos, ni reproducirse. Ninguna otra especie de insecto, ni siquiera las hormigas legionarias, tiene un ejército permanente tan especializado ni dependiente. Como tampoco ningún otro orden de insectos exhibe un arsenal tan diverso de armas químicas y de sistemas de proyección de las mismas.

Los termes evolucionaron, hace unos 150 millones de años, a partir de una estirpe ancestral parecida a las actuales cucarachas. Hoy hay más de 2000 especies de termes en todo el mundo, el 95 por ciento de ellas en los trópicos del Viejo y del Nuevo Mundo. Constituyen el orden de insectos denominado Isópteros ("de alas iguales"), que son los insectos sociales más primitivos.

Los termes, a los que con frecuencia se llama incorrectamente "termitas" y "hormigas blancas", evolucionaron de forma distinta de las hormigas, abejas y avispas sociales del orden Himenópteros. La organización social de los termes se desarrolló por la necesidad de compartir el alimento, hasta lograr el intercambio de bacterias y protozoos simbiontes necesarios para digerir la celulosa. Presentan sociedades rígidamente estructuradas, en las que individuos especializados morfológicamente desempeñan tareas específicas: el rey y la reina se reproducen; los obreros buscan alimento, construyen el refugio y cuidan de las crías; las parejas aladas fecundas salen volando para instaurar nuevas colonias, y los soldados se encargan de la defensa. La comunicación entre los individuos de la colonia se basa en el intercambio de señales químicas; para ello emplean el olfato (por medio de las antenas) o el gusto (por quimiorrecepción de contacto). La construcción, la búsqueda del alimento, el reconocimiento de un compañero de termitero, el seguimiento de pistas, la alarma y la defensa son todos fenómenos en los que intervienen señales químicas específicas.

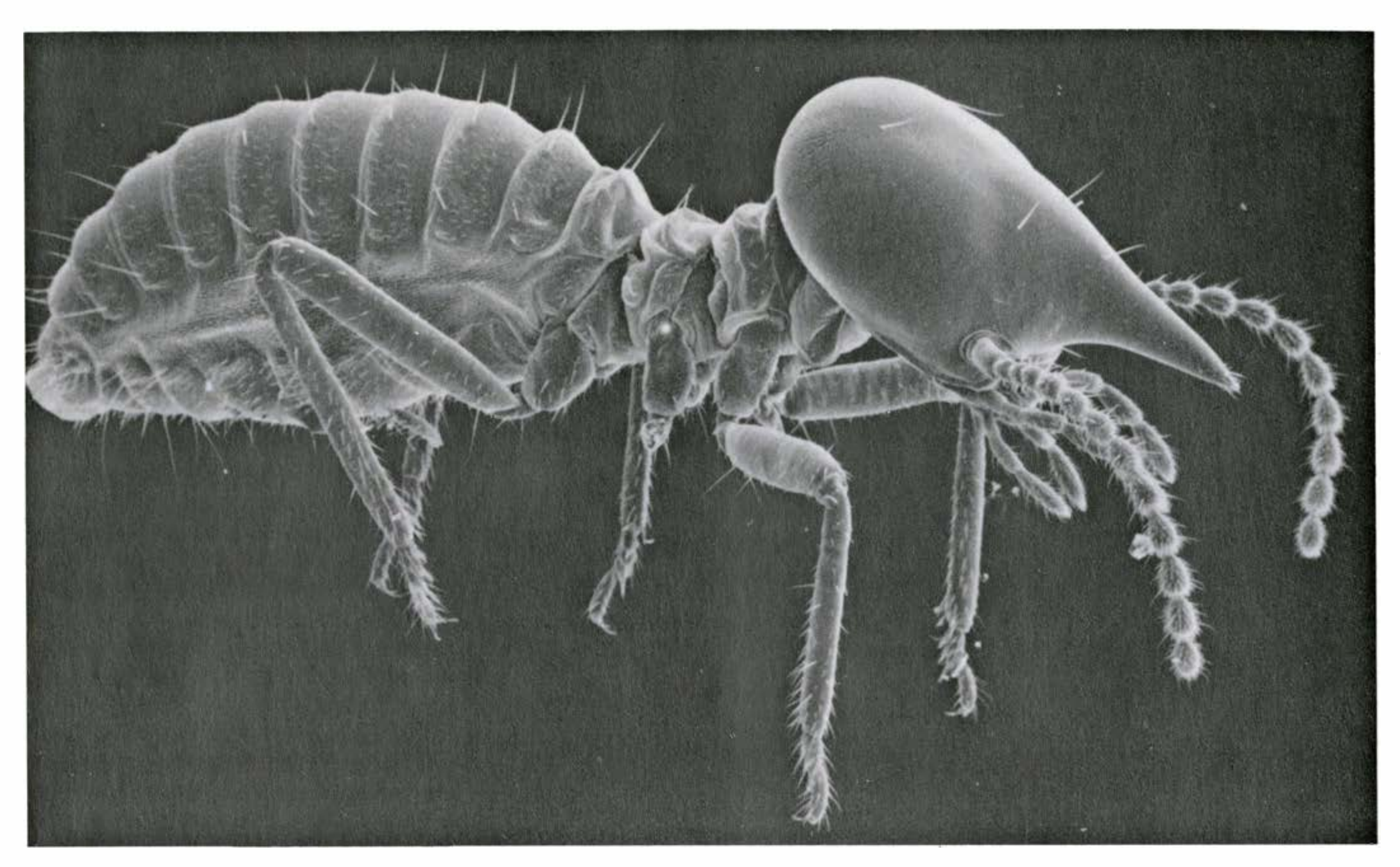
Los termes soldados pueden ser machos estériles o hembras estériles: su cabeza difiere, en forma y tamaño, de la de los termes obreros. Dentro de las actividades de una colonia de termes, a los soldados se les encomienda la protección frente a intrusos. Misión que cumplen ininterrumpidamente, desde la madurez hasta la muerte. La muerte de un soldado representa para la colonia lo que la pérdida de células epidérmicas para el cuerpo humano: aumenta la probabilidad de transferencia genética a la siguiente generación. Los termes soldados son, pues, armas vivas y pueden desempeñar su papel defensivo mediante tácticas que incluyen la mordedura, las dentelladas, el taponamiento de agujeros, la emisión de líquidos, a chorro o por exudación, el embadurnamiento e incluso la defecación autodestructora.

Me extenderé aquí en el comentario de algunos termes soldados representantes tanto de los termes "inferiores" como de los "superiores". Destacaré en particular aquellos géneros de ambos grupos que han complementado sus armas mecánicas (mandíbulas de gran tamaño) con las químicas. A los depredadores intrusos se les puede aplicar muchos tipos de secreción desagradable y de muy diversas maneras. También mostraré de qué modo el análisis de la composición química específica de estas secreciones ayuda a establecer las líneas evolutivas que han seguido las diferentes estirpes de termes, en algunos casos durante 70 millones de años.

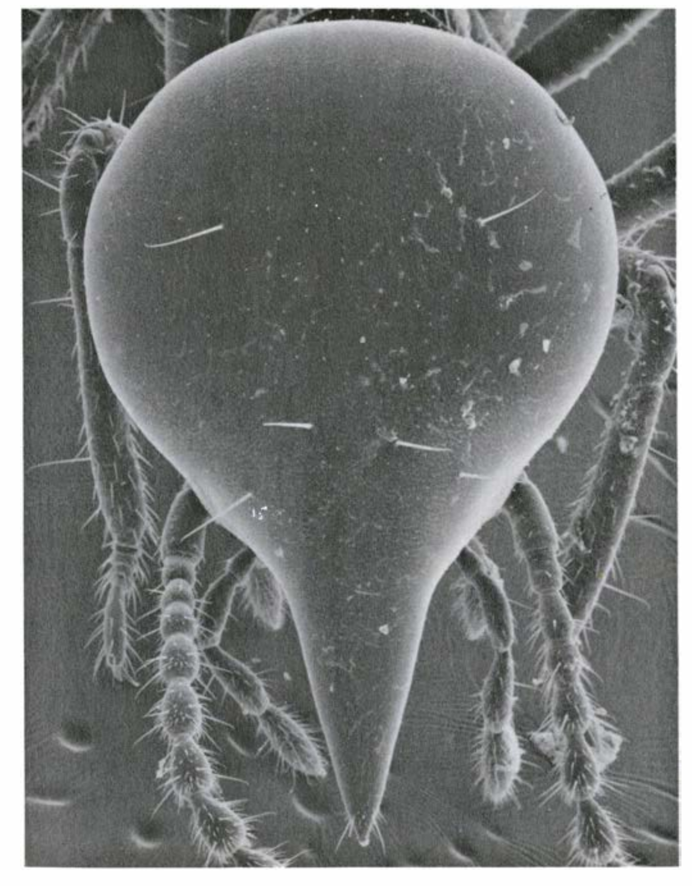
Aunque el arsenal de los termes soldados puede subdividirse en armas mecánicas y armas químicas, el abanico completo de tácticas defensivas es todavía más amplio. Por ejemplo, los soldados de algunas especies perforan la cutícula de un insecto intruso con sus potentes mandíbulas y luego untan la herida con una secreción anticoagulante de su glándula frontal. Ni la herida ni la secreción aplicada serían por sí mismas fatales para el intruso, pero la combinación de una cutícula acuchillada y de un compuesto químico que evita la coagulación, o que es tóxico si penetra en el organismo, supone que el insecto herido acabará por sucumbir.

Al menos otras dos tácticas defensivas no dependen de las mandíbulas ni de secreciones. La cabeza del soldado de *Cryptotermes* tiene forma parecida a un tapón de desagüe; la reacción de los soldados frente a una alarma es dispersarse hacia las estrechas entradas de la colonia y, literalmente, sellar, cual tapones, los agujeros cilíndricos con su cabeza. Las colonias de *Anoplotermes* carecen incluso de una casta de soldados, pero los obreros de la colonia poseen un abdomen rodeado por un músculo constrictor especializado. Cuando el obrero se enfrenta a un intruso, contrae el músculo, con lo que revienta su abdomen y baña al atacante con una mezcla de heces y otros contenidos intestinales.

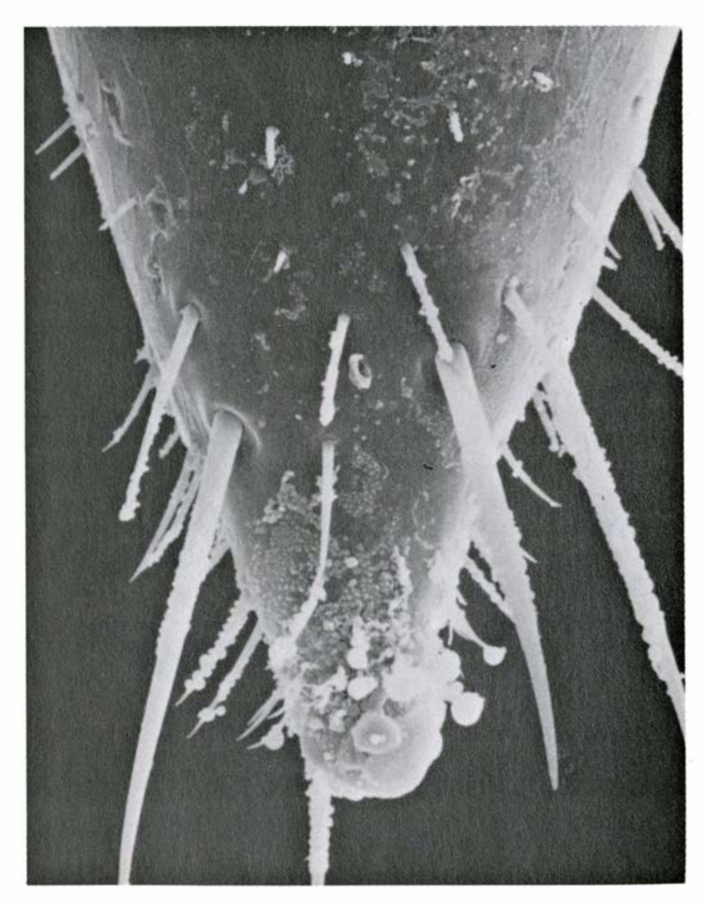
En los últimos años, lo que más ha atraído el interés de los entomólogos es



1. TERMES SOLDADO de la especie *Nasutitermes corniger*, aumentado 50 veces en esta micrografía electrónica de barrido. *N. corniger* es una de las más de 500 especies de la subfamilia de termes superiores, los Nasutitermitinos, que recibe este nombre por el naso (u hocico) alargado que surge de la parte anterior de la cabeza de los soldados. Con esta trompa, semejante a un bazuca, los termes nasudos disparan una sustancia viscosa sobre los intrusos.



2. PARTE SUPERIOR DE LA CABEZA de un soldado de la misma especie de termes, en una micrografía electrónica de barrido, aumentada 70 veces.



3. EXTREMO DEL HOCICO de un soldado de la misma especie de termes, aumentado 1100 veces en esta micrografía electrónica, también de barrido.

la notable variedad de secreciones químicas de los termes. Dos investigadores han estudiado la morfología y la estructura glandular de las armas químicas de los termes soldados: André Quennedey, de la Universidad de Dijon, y Jean Deligne, de la Universidad Libre de Bruselas. Reconocen tres métodos principales de defensa química. El primero consiste en morder, con la introducción simultánea de un material aceitoso o tóxico en la herida. El segundo consiste en untar, aplicando un veneno de contacto a la cutícula del agresor mediante un labro (o labio superior) engrosado, parecido a una brocha de pintor. El tercer método es la emisión de chorros de visco: el soldado rocía al agresor con un agente irritante y viscoso que lo enmaraña.

El primero de los tres métodos se ha desarrollado independientemente varias veces en la familia de los termes superiores, los Termítidos, y también en una de las seis familias de termes inferiores, los Rinotermítidos, que viven bajo tierra. Más de 40 especies "mordedoras-inyectoras" se han estudiado ya en mi laboratorio de la Universidad estatal de Nueva York en Stony Brook y por investigadores de la Unidad de Entomología química de la Universidad de Southampton. La consideración de tres géneros mordedores-inyectores ilustrará parte de la variedad química y morfológica que se ha desarrollado en estas estirpes.

La defensa de la colonia en uno de estos géneros de termes superiores, *Macrotermes*, es tarea de una casta de soldados constituida por hembras estériles con dos tipos de miembros. Los soldados pequeños de *Macrotermes* escoltan a los grandes obreros macho estériles mientras éstos recolectan alimento y reparan el termitero. Los soldados grandes guardan la colmena interior, en la que se guarecen el rey, la reina y las larvas. En ambos tipos de soldados, la glándula frontal de la cabeza secreta una sustancia aceitosa y la exuda a través de la fontanela, una abertura glandular de la cutícula cefálica.

La secreción es la misma, desde el punto de vista químico, en ambos tipos de soldados, pero los soldados grandes llevan un volumen 500 veces superior al que portan los pequeños; de hecho, el peso de tal secreción supone casi el 10 por ciento del peso seco de los soldados grandes. La secreción es un hidrocarburo similar al aceite mineral, o parafina: está constituido por alcanos y alquenos de cadena larga [*véase la figura 5*]. Las cadenas, compuestas de unidades acetato de dos carbonos unidas entre sí, tienen de 21 a 35 átomos de carbono de longitud; la composición exacta de la molécula varía con la especie de termes y la ubicación de la colonia.

Los invasores más comunes de los termiteros son las hormigas. Cuando los soldados de *Macrotermes* luchan con una hormiga, muerden furiosamente con sus mandíbulas, intentando arrancar las patas del intruso. El calor de la actividad hace que la secreción cérea de la glándula frontal se transforme en un aceite líquido que rezuma por la fontanela, resbala por el rostro (o frente) del soldado y se extiende sobre el labro, situado sobre las mandíbulas. De este modo, los mordiscos que dan en el blanco aplican una buena capa de aceite a la cutícula acuchillada de la hormiga. En ausencia de perforación, el hidrocarburo no tóxico es inofensivo. Sin embargo, cuando se aplica a la cutícula herida, parece reblandecer la zona perforada; de ahí que la aplicación, además de interferir la coagulación de la hemolinfa de la hormiga, impide la resclerotización, o reparación natural, de las heridas de la cutícula.



4. SOLDADOS DE OTRA ESPECIE (*Nasutitermes kempae*, que se encuentra en Africa) se arraciman alrededor de la entrada a un conducto de alimentación, abierto para obtener la fotografía. El hocico de cuatro de los soldados apunta hacia la cámara. Esta orientación no es accidental. Aunque los termes son ciegos, los soldados se han orientado en respuesta a las corrientes de aire provocadas por el movimiento del fotógrafo, en este caso el propio autor.

El segundo género mordedor-inyector de los termes superiores es *Cubitermes*, uno de los géneros comedores de suelo que se encuentran en Africa. Los soldados poseen mandíbulas en forma de sable y una cabeza cuadrada y de color anaranjado; también estos animales secretan un hidrocarburo. Sin embargo, la secreción de *Cubitermes* no es una molécula simple de cadena lineal, sino que pertenece a la familia de los terpenos [*véase la figura 5*]. Las moléculas de terpenos se agrupan según el número de unidades pentacarbonadas de isoprenos que poseen. Así, cuando tienen dos unidades isoprénicas (10 carbonos) se les denomina monoterpenos, cuando tienen cuatro unidades (20 carbonos) diterpenos, y así sucesivamente. Los diterpenos aparecen en la secreción de muy pocos insectos; en realidad, y excepción hecha de algunos termes superiores, no se ha demostrado que haya insectos capaces de sintetizarlos. Sin embargo, los soldados de *Cubitermes* figuran en el negocio de fabricación a gran escala de diterpenos. Como colaboradores postdoctorales de Jerrold Meinwald, de la Universidad de Cornell, David Wiemer y yo mismo identificamos tres nuevos diterpenos secretados por este género de termes. De los más de 16 diterpenos que se encuentran en siete especies de *Cubitermes*, en la actualidad sé de al menos cinco que son exclusivos de estos termes. No se conocen otros invertebrados, vertebrados o plantas que sinteticen esos compuestos.

El tercer género de termes mordedores-inyectores es *Armitermes*, un termes comedor de suelo que se encuentra en América Central y del Sur. Los soldados de este género tienen mandíbulas aguzadas, parecidas a un par de pinzas, y poseen una protrusión en forma de gollete sobre la frente. Con el gollete vierten gotitas de una secreción aceitosa a las heridas perforantes que infligen a los intrusos. Gracias a la diligente recolección que efectuó Margaret S. Collins, de la Universidad de Howard y colaboradora mía, en la pluviselva de Guyana, he tenido acceso a soldados de varias especies de *Armitermes* para analizar sus secreciones.

Las secreciones de *Armitermes* eran ácidos grasos modificados, de 22 a 36 átomos de carbono por cadena molecular. Un extremo de la cadena portaba un grupo de ácido carboxílico, y el otro extremo un grupo hidroxilo. Cadenas similares se encuentran en la lanolina de la lana, pero lo insólito de estas secreciones es que los dos extremos de la cadena estaban unidos mediante un enlace éster, formando un bucle molecular. Bucles similares con menos átomos de carbono, denominados macrólidos, se encuentran entre las sustancias que actúan como atrayentes sexuales de los mamíferos, y entre las que tapizan los nidos de las abejas solitarias. Sin embargo, que yo sepa, los macrólidos de 22 a 36 átomos de carbono son exclusivos de *Armitermes* y de un género mordedor-inyector emparentado de termes superiores, *Rhynchotermes*. Barbara L. Thorne, de la Universidad de Harvard, y James F. A. Traniello, de la Universidad de Boston, han descubierto que cuando la secreción se aplica a la herida perforada de un intruso actúa como una toxina, y no sólo como inhibidor de la curación.

5. TIPOS DE MOLECULAS que se encuentran en las secreciones de los termes, representadas por su fórmula estructural. Los termes del género *Macrotermes* segregan alcanos: hidrocarburos de cadena lineal (*arriba*). Los termes de los géneros *Nasutitermes* y *Cubitermes* segregan hidrocarburos terpénicos, más complejos (*centro*). Los *Nasutitermes*, más avanzados, segregan terpenos todavía más complejos.

El segundo de los tres métodos principales de defensa química de los termes soldados, según Quennedey y Deligne, contempla el embadurnado del intruso con una secreción tóxica de lípidos químicamente reactivos derivados de ácidos grasos. Los termes soldados de la familia Rinotermítidos, termes inferiores, se hallan adaptados morfológicamente para este modo de defensa. Su labro se ha ensanchado hasta formar una especie de brocha de pintor cerdosa que el soldado presiona contra el intruso, extendiendo el corrosivo veneno de contacto, soluble en grasas, sobre su cutícula. A diferencia de los termes superiores mordedores-inyectores, estas especies no almacenan sus secreciones sólo en la cabeza, sino que han desarrollado un gran depósito de reserva en el abdomen. En algunos soldados, la capacidad conjunta de estos depósitos sobrepasa el 35 por ciento del peso seco de estos animales.

Quennedey y sus colegas, de Dijon, describieron en 1973, por primera vez, el veneno de un género africano, *Schedorhinotermes*, como una vinilcetona. Por esa misma época, Ján Vrkoč y sus colegas, de la Academia checoslovaca de Ciencias, informaron que el veneno del género pantropical *Prorhinotermes* era un nitroalqueno. Desde entonces mis colegas y yo hemos confirmado y ampliado ambas identificaciones en especies emparentadas. Hemos informado asimismo de un tercer tipo de venenos, los beta-cetaldehídos, secretados por los soldados de dos rinotermítidos del Nuevo Mundo: los géneros *Rhinotermes* y *Acorhinotermes*.

Aunque cada uno de estos venenos de contacto tiene una estructura molecular distinta, los tres comparten dos características clave. En primer lugar,

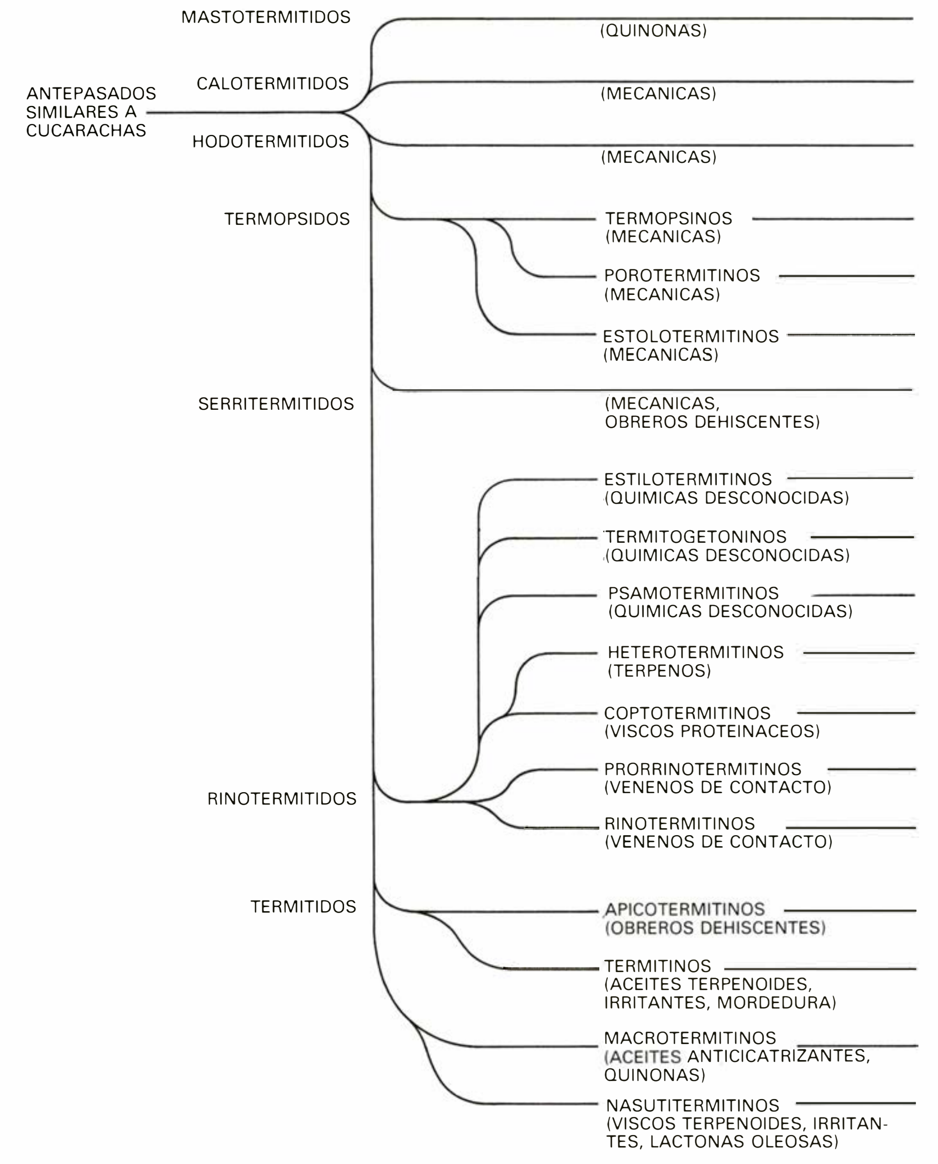
son derivados de moléculas de ácidos grasos de 14 a 16 átomos de carbono, y todas adquieren la forma de largas cadenas carbonadas que son liposolubles. En segundo lugar, las tres poseen, en un extremo de la cadena, un grupo de átomos electrófilo (que muestra avidez por los electrones) químicamente reactivo. Podrían compararse a una flecha envenenada. El astil soluble en grasas de la "flecha" facilita su paso a través de la cutícula cérea del intruso. Una vez dentro de la cutícula, la punta emponzoñada de la "flecha", el grupo electrófilo, produce lesiones químicas internas. Tales grupos electrófilos constituyen una característica común en las secreciones defensivas naturales. Se encuentran en las secreciones de muchos hongos y plantas superiores. Pueden actuar como antibióticos, inhibidores tumorales y repelentes de insectos.

¿De qué modo evitan estos termes envenenarse a sí mismos con las toxinas que secretan? Deben haber desarrollado adaptaciones bioquímicas que les permiten detoxificar sus propios electrófilos; en ausencia de algún método de detoxificación, los termes soldados no podrían sobrevivir a la síntesis ni al almacenamiento de sus armas químicas, y los termes obreros no soportarían su despliegue. No obstante, hasta hace poco nadie había determinado la base bioquímica de la inmunidad de los termes. Se especulaba que el glutatión (GSH), un tripéptido que contiene azufre, y un grupo de enzimas detoxificantes, las S-transferasas del glutatión, se hallaban involucrados en el proceso. Las S-transferasas del GSH hacen que los electrófilos sean más solubles en agua y menos reactivos químicamente, con lo que facilitan su excreción. Se sabe que tanto los vertebrados como los invertebrados, así como las plantas, emplean las S-transferasas del GSH para desembarazar de sustancias tóxicas su sistema. En los termes se encuentran también niveles moderados de GSH y de S-transferasa del GSH.

Sin embargo, uno de mis alumnos, Stephen Spanton, y yo descubrimos que los obreros del termes de Florida *Prorhinotermes simplex*, cuyos soldados sintetizan nitroalquenos tóxicos, y los obreros del termes africano *Schedorhinotermes lamanianus*, cuyos soldados sintetizan vinilcetonas tóxicas, detoxifican las secreciones de sus soldados de otra manera. Reducen los dobles enlaces, pobres en electrones, de las moléculas, convirtiendo los nitroalquenos en nitroalcanos y las vinilcetonas en etilcetonas saturadas. Los productos reducidos poseen sólo la décima parte de la toxicidad de las secreciones no reducidas.

Encontramos que la detoxificación inicial se conseguía por la acción de un enzima que, específicamente, reduce el alqueno electrófilo de esta especie y que requiere NADPH, un cofactor nucleótido reducido, como fuente de hidrógeno. En los obreros de *Schedorhinotermes*, el 50 por ciento de la vinilcetona se convirtió en etilcetona en menos de una hora. Toda la etilcetona se catabolizó (descompuso) luego en acetato en un par de días. Los obreros de *Prorhinotermes* reducían enzimáticamente la toxina de sus soldados de un modo similar. Sin embargo, al exponer estos obreros a la toxina de soldados de *Schedorhinotermes*, no toleraban dosis del calibre de las que soportaban cuando se trataba de la toxina de sus propios soldados.

Con excepción de estas dos especies de termes, la ruta del catabolismo de reducción para detoxificar venenos electrófilos es, por lo que se ve, una rareza. Ello puede deberse a que los termes se hallan sometidos a presión evolutiva para el desarrollo de rutas alternativas de detoxificación con el fin de conservar nitrógeno para la síntesis de proteínas: su alimento, rico en celulosa, es pobre en nitrógeno. Los termes practican estrategias de conservación de nitrógeno tales como la alimentación selectiva y el canibalismo.



6. SIETE FAMILIAS DE TERMES surgieron (*izquierda*) de una estirpe ancestral de insectos parecidos a cucarachas. Las seis familias conocidas colectivamente como termes inferiores dependen principalmente de defensas mecánicas: las poderosas mandíbulas de sus soldados. Sin embargo, los Rinotermítidos comprenden siete subfamilias, de las que seis poseen soldados que utilizan defensas químicas, ya sea junto con la mordedura o sin ella. La séptima familia de termes, los Termítidos (*abajo*), incluye las cuatro subfamilias de termes superiores. Una subfamilia, los Apicotermitinos, no tienen casta de soldados. La defensa de la colonia está a cargo de trabajadores dehiscentes: explotan y rocían al intruso con el contenido intestinal. Los Nasutitermitinos se encuentran en los trópicos del Viejo y del Nuevo Mundo. Parece claro que evolucionaron hace unos 70 millones de años. Las defensas de cada grupo se indican entre paréntesis.

PRORHINOTERMES

H_3C … NITROALQUENO

RHINOTERMES (MINOR)

H_2C … BETA-CETALDEHIDO

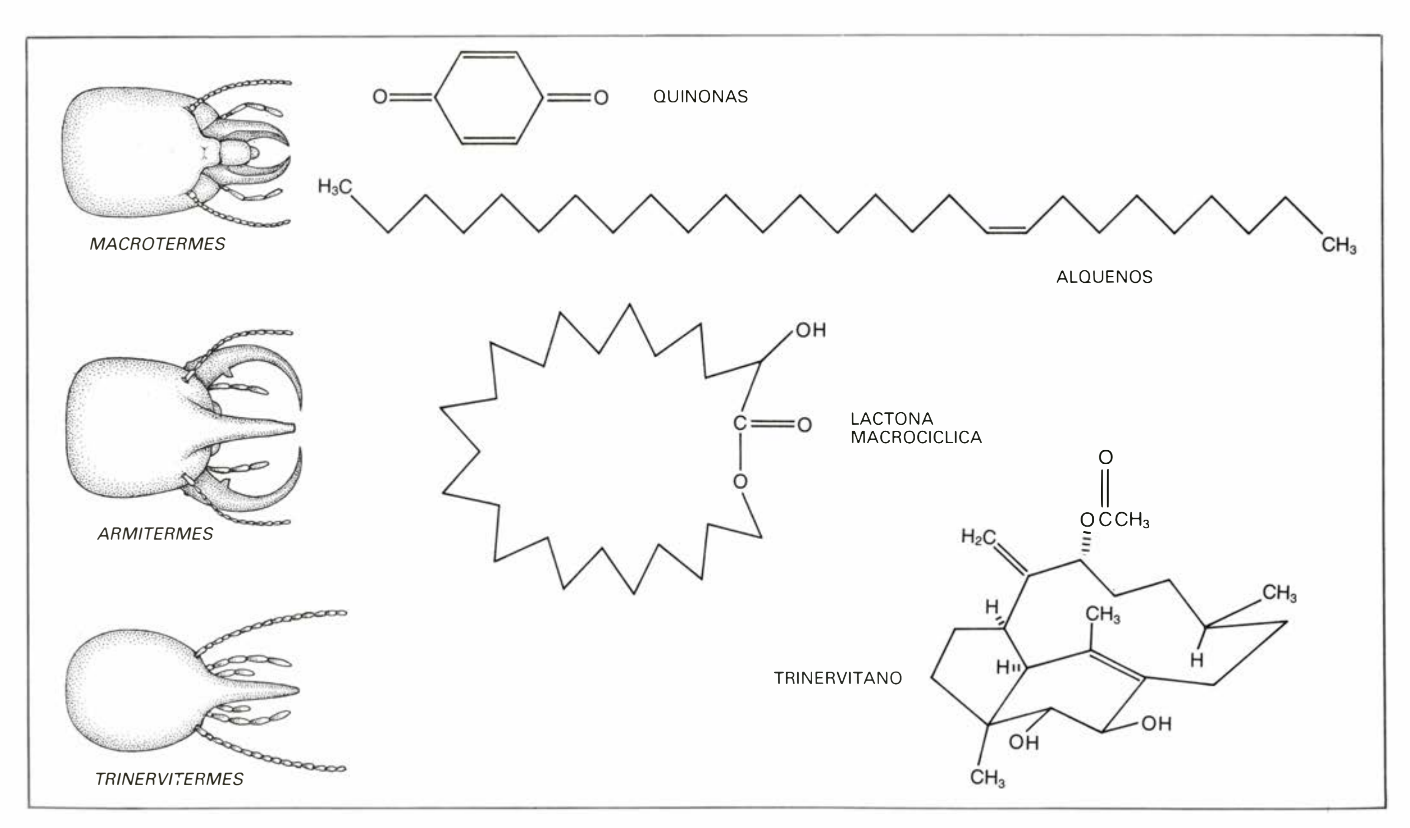
SCHEDORHINOTERMES

H_2C … CH_2 VINILCETONAS

ACORHINOTERMES

H_3C … BETA-CETALDEHIDO

7. MODIFICACIONES CEFALICAS de los soldados de cuatro géneros de termes de la familia Rinotermítidos; se indican desde arriba, a la izquierda. La estructura molecular de las secreciones defensivas que cada soldado aplica a un invasor de su colonia se encuentra a la derecha. Los hidrocarburos de cadena larga secretados por los cuatro géneros van desde un nitroalqueno simple hasta una vinilcetona y a un beta-cetaldehído, de mayor reactividad.



8. MODIFICACIONES CEFALICAS similares de los soldados en tres géneros de termes superiores. Abajo se ilustra la modificación típica del género nasudo *Trinervitermes*. Las secreciones de *Macrotermes* pueden ser quinonas, alcanos o alquenos simples. El género de nasudos *Armitermes* secreta bucles moleculares denominados macrólidos. La secreción de *Trinervitermes* es una mezcla de diterpenos, única de la que hay constancia en la naturaleza.

Cathy J. Potrikus y John A. Breznak, de la Universidad estatal de Michigan, han demostrado que las bacterias simbiontes de los termes pueden fijar nitrógeno atmosférico y reciclar el ácido úrico, un compuesto nitrogenado; ambas actividades contribuyen de forma importante al suministro de nitrógeno a los termes. Barbara L. Bentley, de la Universidad estatal de Nueva York en Stony Brook, y yo hemos demostrado que los géneros de *Nasutitermes* de Costa Rica fijan suficiente nitrógeno molecular para duplicar su reserva de nitrógeno en menos de seis meses. No obstante, los termes floridanos y africanos resistentes a los venenos electrófilos parecen evitar la pérdida de nitrógeno durante la detoxificación reciclando, en lugar de excretando, tanto los aminoácidos ricos en nitrógeno empleados en la detoxificación como la energía almacenada en la cadena de carbonos de la secreción defensiva.

El último método importante de defensa química en los termes, y en un sentido evolutivo el más avanzado de los tres, requiere todavía otra modificación de la anatomía de los soldados. En lugar de embadurnar a un intruso con una toxina, los soldados de unas 500 especies de una abundante subfamilia de termes pantropicales emplean una frente modificada dotada de nariz, o naso, es decir, un tubo hociquiforme. Esta subfamilia es la de los Nasutitermitinos, así llamada por el naso. Cuando un termitero de nasudos o narigudos es atacado, los soldados se desplazan desde el interior al punto de invasión y expulsan por su naso una secreción pegajosa e irritante que enmaraña al intruso. Los obreros se unen a la acción, mordiendo al invasor con sus aguzadas mandíbulas e intentando dañarle sus patas. Hormigas, arañas y otros animales insectívoros, incluso los osos hormigueros, suelen evitar la maloliente, desagradable y potencialmente letal barrera de los soldados nasudos.

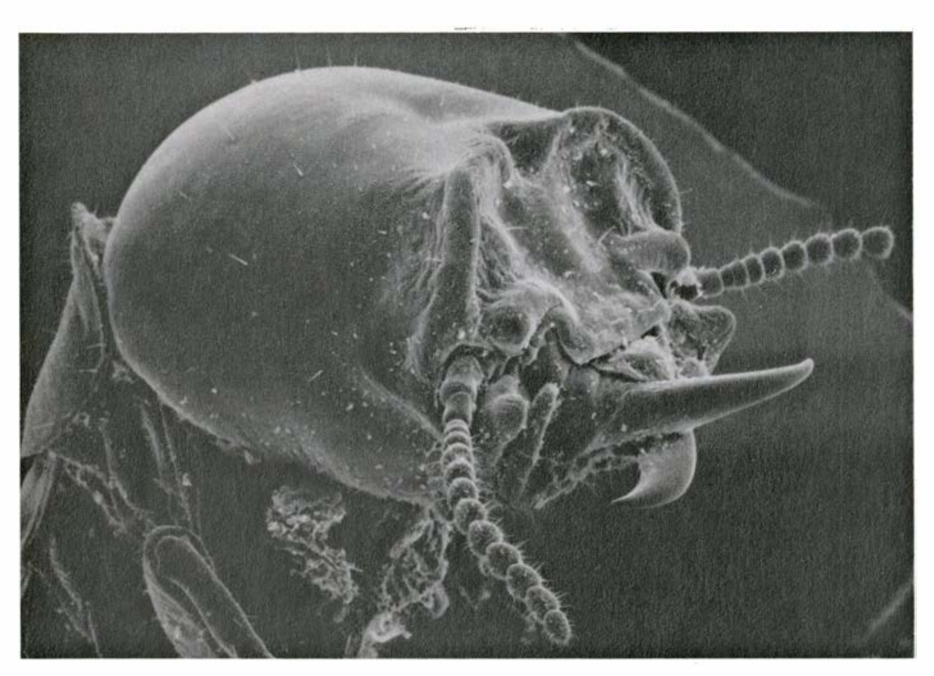
Desde hace algún tiempo se sabía que la secreción defensiva viscosa de los nasudos era análoga a la savia de pino: una mezcla de hidrocarburos monoterpénicos (el solvente) y de otros isoprenoides de peso molecular más alto (la resina). Pero la secreción de los termes difiere de la savia de pino en que no contiene ninguno de los ácidos resínicos usuales. En 1974, cuando mis colegas y yo emprendimos el estudio de este visco de los narigudos, teníamos planteadas muchas preguntas. ¿Cuáles eran exactamente los compuestos pesados? ¿Eran simplemente pegajosos o eran también tóxicos? ¿Procedían los soldados a la biosíntesis activa de sustancias, o bien se adquirían éstas a partir del alimento? Los termes narigudos constituyen la subfamilia de termes más numerosa y diversa; su distribución es mundial; ¿cómo varía la composición del visco de una especie a otra y de un lugar a otro? Hoy, nueve años después, tenemos ya la respuesta a algunas de las cuestiones.

Nuestro primer éxito llegó con el aislamiento y purificación de varios componentes del pegamento y con la interpretación de su estructura. Ello requirió la combinación de mis investigaciones en Africa con los esfuerzos de Jon C. Clardy y sus colegas, de la Universidad estatal de Iowa, y de Koji Nakanishi y sus colegas, de la Universidad de Columbia. Los primeros componentes del visco se identificaron a partir de miles de soldados nasudos del género africano *Trinervitermes*, de termes recolectores de hierba, de ahí que se les denominara trinervitanos. Se trataba de moléculas diterpénicas, que constaban de tres anillos fusionados con su esqueleto de átomos de carbono dispuesto en forma de cúpula. El esqueleto de los carbonos no se asemejaba a ninguna otra estructura molecular natural conocida.

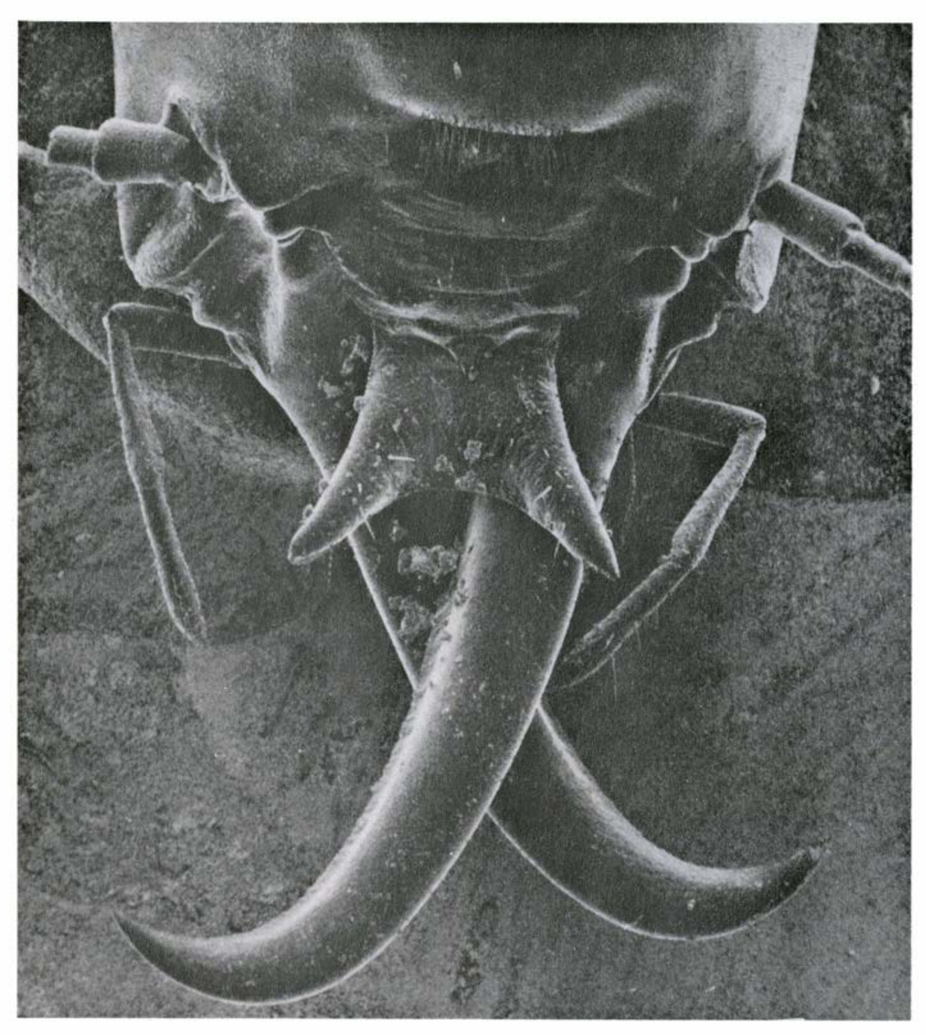
Desde entonces, nosotros y otros investigadores hemos aislado un diterpeno bicíclico (de dos anillos) y tres tetracíclicos (de cuatro anillos), que se conocen, respectivamente, por los nombres de secotrinervitano, quempano, ripertano y longipano. Junto a otras 60 sustancias derivadas, son secretados por los soldados de otros géneros de termes narigudos. Todos muestran las mismas características estructurales generales. Cada molécula tiene el mismo progenitor biosintético: el cembreno-A, un diterpeno monocíclico. Todas son cupuliformes; de la superficie convexa se proyectan grupos que contienen oxígeno y que atraen el agua, y, en la superficie cóncava, se halla una región hidrorrepelente. La defensa contra los intrusos se consigue, por tanto, con una sustancia bastante parecida a la savia de pino: una solución viscosa de diterpenos mezclados asociada a solventes monoterpénicos. La mezcla moja muy bien la cutícula de los insectos y otros artrópodos, que normalmente es hidrorrepelente.

Hemos comprobado asimismo que las secreciones varían de un soldado a otro en una misma colonia, entre soldados de poblaciones conectadas de la misma especie, entre poblaciones geográficamente aisladas de la misma especie y entre especies del mismo género. Este tipo de variación química, que es más común en el reino vegetal que en el animal, puede ser de utilidad en los estudios taxonómicos.

¿Cuál es, exactamente, el efecto que ejerce el visco del nasudo? Thomas Eisner y sus colegas, de Cornell, han demostrado en una serie de elegantes experimentos de laboratorio que funciona como un agente enmarañador, como un irritante (propiedad que favorece su dispersión mientras el intruso rociado se limpia) y como veneno tópico. Quedaba por determinar si



9. CABEZA ENCAJADA, adaptación del soldado de *Cryptotermes*. Cuando estos soldados son alarmados por la presencia de un intruso, taponan la entrada a la colonia con sus cabezas, mostrando una formación de mandíbulas para disuadir a otros invasores. Estos termes carecen de defensas químicas.

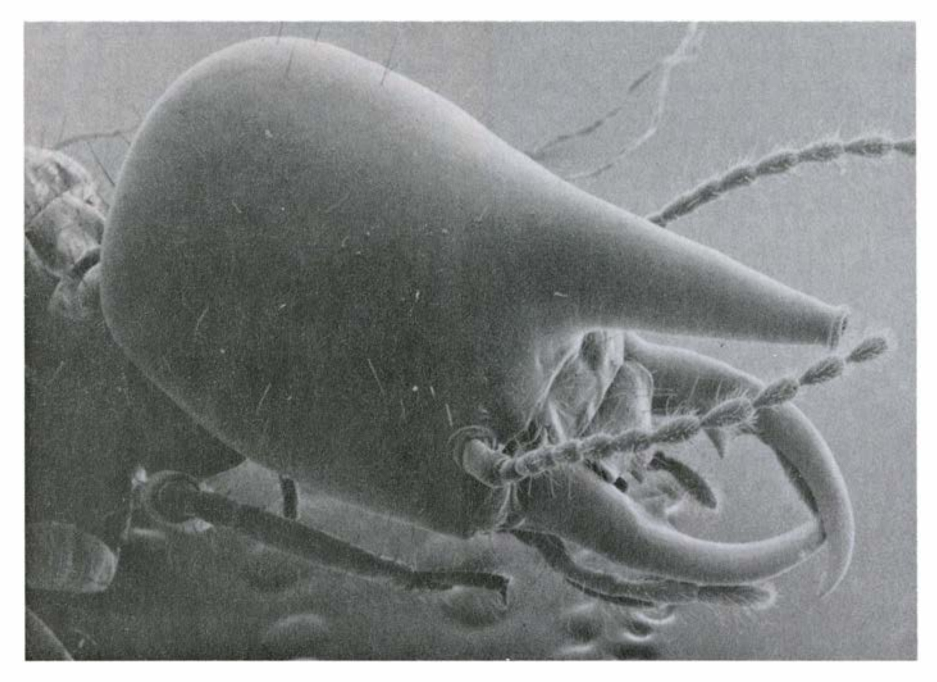


10. MANDIBULAS EN FORMA DE SABLE de un soldado de *Cubitermes*; pueden acuchillar la cutícula de una hormiga invasora mientras una secreción de la glándula frontal del soldado se extiende sobre la herida. La secreción parece inhibir la coagulación de la sangre y la reparación de la cutícula del intruso.

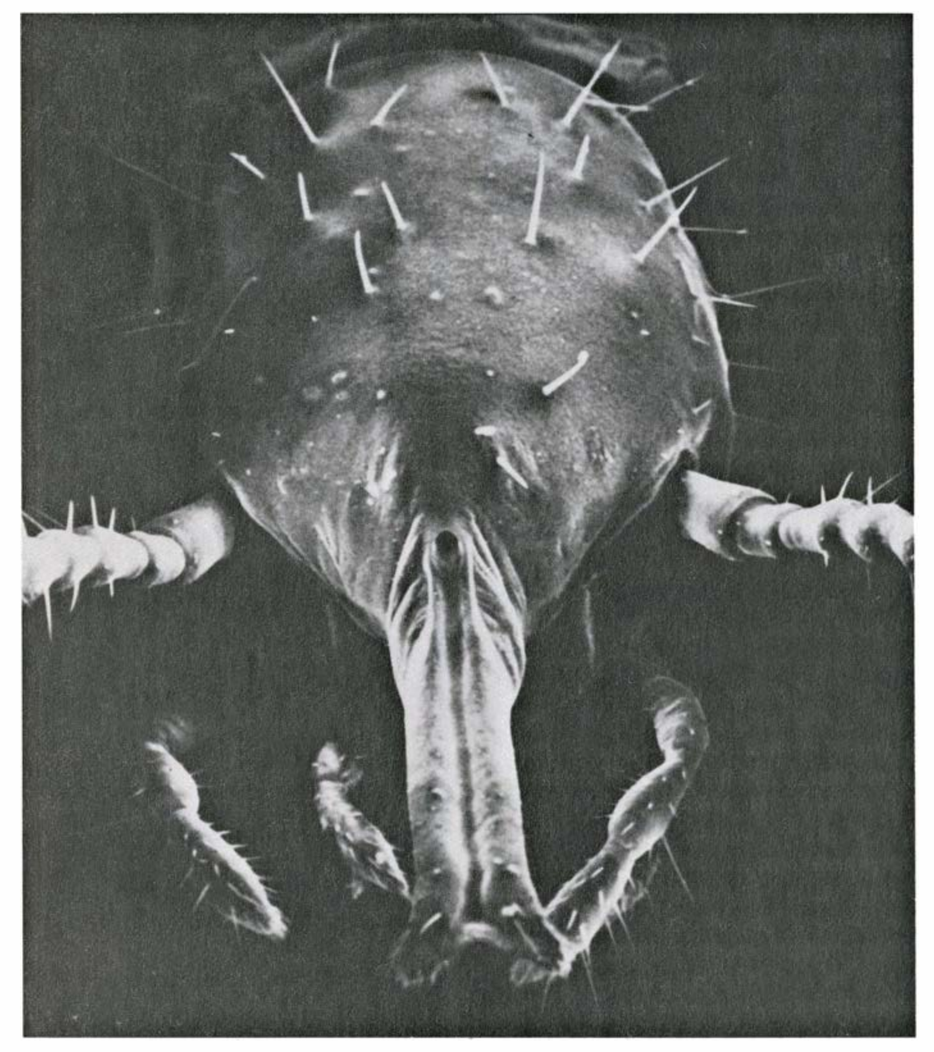
los soldados sintetizaban la cola o bien la obtenían del alimento, cuestión que mis colegas y yo decidimos resolver mediante el empleo de trazadores radiactivos. No se puede incluir un precursor marcado en la dieta de los termes soldados, porque el único nutrimento que éstos aceptan es el fluido que los obreros les regurgitan. Atacamos el problema valiéndonos de una micropipeta, con la que inyectábamos moléculas precursoras marcadas radiactivamente (acetato sódico o mevalonato sódico) directamente en el abdomen de los soldados. Debido a la elevada flexibilidad de la pared abdominal de los termes, podíamos inyectar cerca de medio microlitro en cada individuo, con lo que se le duplicaba, aproximadamente, el tamaño del abdomen.

Los soldados sobrevivían varias horas a la inyección. Para entonces, los precursores se habían incorporado a los monoterpenos y diterpenos almacenados en la cabeza de los soldados. Cuando inyectamos en obreros (que no secretan diterpenos) los mismos precursores marcados, no se encontraron precursores en sus compuestos terpenoideos. Después de una cuidadosa purificación de cada componente llegamos a la conclusión de que los soldados nasudos sintetizaban efectivamente sus secreciones defensivas *de novo*. Su capacidad para hacerlo los sitúa aparte de todos los demás insectos (quizá con la excepción de algunos cóccidos). Ningún otro insecto puede sintetizar terpenos de tamaño superior a un sesquiterpeno de 15 átomos de carbono.

En otro experimento se demostró que los diterpenos secretados por un soldado nasudo no resultaban afectados por la dieta del soldado. El método consistió en criar dos poblaciones químicamente distintas de una única especie de *Trinervitermes*, partiendo de parejas separadas de reproductores alados, machos y hembras, después de su vuelo nupcial. A ambas parejas y a las colonias a las que dieron origen se les suministró el mismo alimento: hierba seca de uno de los lugares de recolección. A pesar de su dieta común, los soldados que llegaron a la madurez seis meses después en cada nueva colonia



11. SOLDADO NASUDO DE TERMES del género *Armitermes*, que se defiende químicamente de forma similar al anterior. Después de perforar la cutícula de un atacante con sus mandíbulas ganchudas emplea su hocico para aplicar una secreción grasa a la zona herida. La mezcla es tóxica si penetra en la herida.



12. LABIO SUPERIOR LARGO de un soldado del género *Rhinotermes*. Lo utiliza para embadurnar la cutícula de un intruso con un cetaldehído tóxico que almacena en su cabeza y abdomen. El veneno, que es soluble en grasas, penetra en la cutícula cérea del intruso e interfiere con los procesos metabólicos.

produjeron diterpenos químicamente idénticos a los de sus colonias paternas.

Los análisis químicos de las secreciones de los soldados narigudos han iluminado un aspecto particular de la evolución de los termes. Ha sido habitual considerar que la subfamilia pantropical de nasudos Nasutitermitinos era un ejemplo capital de evolución paralela. Es decir, que los géneros narigudos de Africa, Asia, Australia y Sudámerica surgieron de termes ancestrales cuyos soldados estaban armados únicamente con mandíbulas. En el decurso de miles de generaciones, los soldados primitivos (que en la actualidad se encuentran sólo en Sudamérica) sufrieron una regresión de sus mandíbulas y una modificación de su frente, que la convirtió en el naso de aspecto de bazuca. Sin embargo, parece demasiada exigencia de un proceso evolutivo paralelo que la capacidad de sintetizar diterpenos defensivos, una hazaña química única de los soldados avanzados de termes nasudos, haya evolucionado también independientemente. En lugar de esta hipótesis, sugiero que en Gondwana occidental existió una población ancestral común de termes elaboradores de diterpenos, antes de que la tectónica de placas dividiera esa masa continental en Protoáfrica y Protosudamérica, en tiempos cretácicos.

Viene a apoyar ese litigio el hecho de que muchos géneros nasudos "primitivos" actuales pueden hacerse encajar en un esquema ancestral en el que las mandíbulas se hacen más pequeñas y el naso se alarga. Sin embargo, el examen químico de las secreciones de los géneros intermedios no muestra ninguno de los cuatro diterpenos únicos que caracterizan las secreciones de los nasudos avanzados. Los narigudos sudamericanos con mandíbulas grandes no pueden proyectar sus secreciones; únicamente pueden hacerlas rezumar en gotitas, y lo que fluye está compuesto por macrólidos derivados de ácidos grasos y mono- y sesquiterpenos. En cambio, los nasudos sudamericanos, más "avanzados", carecen de mandíbulas funcionales, pero pueden expulsar soluciones de diterpenos como sus parientes africanos, asiáticos y australianos. La estructura química tridimensional idéntica de todos los diterpenos de nasudos de todo el mundo apoya la hipótesis de un origen evolutivo común. Así pues, parece ser que lo que se consideraba un excelente ejemplo de evolución paralela debe borrarse de los libros. Parece asimismo probable que estudios ulteriores de la química defensiva de los termes nos deparen nuevas sorpresas.

Decisiones racionales colectivas

El análisis axiomático de los sistemas de votación pone en entredicho la compatibilidad de varias propiedades deseables en métodos ideales. Es inevitable la transacción entre racionalidad, decisividad e igualitarismo

Douglas H. Blair y Robert A. Pollack

¿Será posible idear un sistema de votación que sea a un tiempo racional, decisivo e igualitario? Estudios efectuados por filósofos, especialistas en ciencias políticas y economistas (los autores, entre ellos) hacen pensar que la respuesta haya de ser negativa. En efecto, estas características idealmente exigibles son en realidad contradictorias. Un método de votación puede evitar la arbitrariedad; puede evitar empates y puntos muertos, puede evitar que el valor del voto sea desigual. Pero no puede escapar de los tres escollos a la vez. El análisis de este dilema, que continúa, ha permitido comprender más a fondo los sistemas de votación existentes, y quizá, con el tiempo, nos permita encontrar otros mejores.

El análisis axiomático de los procedimientos racionales de votación lo inició, hace unos 33 años, el economista Kenneth J. Arrow, de la Universidad de Stanford. Arrow propuso cinco axiomas, intuitivamente satisfactorios, que debería satisfacer todo procedimiento de combinación o agregación de preferencias individuales encaminado a sintetizar un dictamen colectivo, y demostró que los únicos procedimientos que pueden obedecer a todos ellos concentran todo el poder en manos de un único individuo. No es posible hallar un método que satisfaga todos los axiomas de Arrow sin ser de carácter dictatorial, y ello, no por menester de ingenio, sino porque ninguno existe. Arrow recibió en 1972 el premio Nobel de Economía, en parte, por este trabajo.

A lo largo de los últimos 15 años, los investigadores han cribado los axiomas de Arrow, buscando en la relajación de sus prescripciones alguna escapatoria a su "teorema de imposibilidad". El problema ha merecido muy amplia atención, porque está íntimamente ligado a cuestiones fundamentales de la filosofía, las ciencias políticas y la economía. Los filósofos, por ejemplo, se topan con él al analizar las consecuencias prácticas del utilitarismo, doctrina ética que postula que la licitud y rectitud de las acciones depende de sus repercusiones en la felicidad del pueblo, lo que exige, por consiguiente, un procedimiento de agregación de preferencias de los individuos. Los estudiosos de ciencia política se lo encuentran al diseñar o evaluar reglas de votación de comisiones o cuerpos legislativos. Los economistas han de planteárselo al analizar sistemas de racionamiento y otros métodos de asignación de recursos que no dependan del juego del mercado. El problema tiene especial importancia en economía normativa, porque, para determinar el alcance apropiado de la intervención gubernamental en los mecanismos de una economía de mercado, es crucial comprender correctamente qué rendimientos potenciales tendrán las distintas alternativas enfrentadas al puro *laissez faire*.

La regla de decisión por mayoría merece ser considerada la primera, entre las distintas formas de agregación de voluntades individuales. Entre sus virtudes cuenta la de ser simple, igualitaria y de larga tradición. La regla de decisión por mayoría, en esencia, es un procedimiento para optar entre pares de candidatos o alternativas. Sin embargo, cuando en la liza entran más de dos candidaturas, la regla de mayoría tropieza con una dificultad ya detectada hace casi 200 años por el Marqués de Condorcet.

La dificultad señalada por Condorcet se conoce hoy por "paradoja del voto". Supongamos que una comisión compuesta por Antonio, Benito y César deba ordenar por preferencia a tres candidatos, *x*, *y*, y *z*. El orden de prelación que les asignaría Antonio es *x*, *y*, *z*. El de Benito es *y*, *z*, *x*, y el de César, *z*, *x*, *y*. La votación por mayoría entre pares de candidatos provoca un ciclo: *x* vence a *y*, *y* vence a *z*,y *z* vence a *x*, en todos los casos, por dos votos contra uno. Este ciclo es el ejemplo más sencillo de paradoja del voto de Condorcet.

Los especialistas en ciencias políticas han detectado muchos casos históricos de ciclos de votación. Por ejemplo, William H. Riker, de la Universidad de Rochester, aduce que la adopción de la 17.ª Enmienda a la Constitución de los Estados Unidos, que establece la elección directa de los senadores, fue retrasada durante diez años mediante maniobras parlamentarias basadas en ciclos de votación entre, por una parte, el *status quo* (que consistía en el nombramiento de los senadores por el cuerpo legislativo de cada Estado) y, por otra, dos variantes de la enmienda.

En cuanto sean factibles más de dos alternativas, se introducirá un nuevo principio para elegir y decidir partiendo de prelaciones por parejas (ordenación entre pares). Las configuraciones de preferencias capaces de inducir la paradoja de Condorcet suscitan dificultades en cada una de las que podríamos llamar "soluciones naturales". El método más simple consistiría en elegir una alternativa que no pudiera ser individual y aisladamente derrotada por ninguna otra (lo que se conoce por "mayoría simple"). Sin embargo, en situaciones como la explicada en la paradoja del voto, no existen alternativas así, pues cada alternativa es derrotada siempre por alguna de las otras.

Un segundo método para proceder a seleccionar una candidatura basándonos en preferencias sobre pares consiste en especificar un programa u orden de procedimiento donde se detalle en qué secuencia deben considerarse los pares de alternativas. Por ejemplo, el orden procesal podría requerir que se votase primero entre *x* e *y*, para luego, en segunda fase, enfrentar al vencedor de estos dos con *z*. Dado este programa, nuestra comisión tripartita del ejemplo comenzaría prefiriendo *x* frente a *y*, para luego, en segunda ronda,

preferir a z y desechar a x. Pero es fácil comprobar que, en cualquiera de los tres órdenes procesales posibles en la situación, es siempre la candidatura que ocupe el último lugar la que sale vencedora: es el orden de procedimiento quien determina el resultado. Los ciclos de votación suscitan no sólo dificultades de carácter estético, sino también sustantivo. Cuando llega a establecerse uno de estos ciclos, la determinación del vencedor definitivo es, en el mejor de los casos, arbitraria (si el orden de procedimiento fue decidido por sorteo) y en el peor, fruto de las maquinaciones de quien tenga que decidirlo.

Surgen nuevas oportunidades para maniobras tácticas si alguno de los electores tiene capacidad para introducir nuevas alternativas, alterando así el orden de procedimiento. Supongamos que (con la misma comisión del ejemplo) z signifique mantener el *status quo*, y que y sea una alternativa incorporada a una moción presentada. Con sólo estas dos posibilidades, y vencería a z, lo que dejaría descontento a César, quien prefiere z antes que y. Sin embargo, si le fuera posible presentar la enmienda x a la moción y, x vencería a y en la primera ronda de la votación, mientras en la segunda x resultaría derrotada por z. De esta forma César conseguiría poner en vigor la alternativa por él más favorecida.

Aun cuando no fuera posible introducir alternativas ni manipular el orden procesal, sí caben todavía beneficios marginales para los electores, si éstos falsean el votar sus verdaderas preferencias. Volvamos al ejemplo en que z entra en último lugar. Si cada miembro de la comisión votase su verdadera preferencia en cada ronda, la alternativa victoriosa, z, sería la menos deseable por Antonio. Supongamos, empero, que Antonio optase por y, y no por x en la votación inicial. Prevalecería entonces y, quien conseguiría vencer a z en la segunda fase. Gracias a esta estratagema, Antonio logra impedir que resulte elegida la alternativa que más le disgusta.

MIEMBROS DE LA COMISION	ORDENACION PREFERIDA	RESULTADOS DE LA VOTACION POR MAYORIA
ANTONIO	x, y, z	z VENCE A x, 2-1; x VENCE A y, 2-1; y VENCE A z, 2-1
BENITO	y, z, x	
CESAR	z, x, y	

1. PARADOJA DEL VOTO que puede producirse al aplicar la regla de la mayoría, cuando electores con preferencias en conflicto hayan de elegir entre más de dos alternativas. Vemos la paradoja ilustrada para un comité de tres miembros que estudia tres candidaturas o alternativas, x, y, z. El resultado es cíclico: x vence a y, y vence a z y z vence a x, en todos los casos por dos votos contra uno. Tales problemas llevaron a Kenneth J. Arrow a proponer cinco axiomas que todos los sistemas de votación deben satisfacer.

Los ciclos en las preferencias colectivas suscitan problemas de arbitrariedad en los resultados y son vulnerables frente a tácticas maniobreras. Estas dificultades se presentan tanto si el procedimiento para ir estableciendo preferencias es el de mayoría como si se utiliza cualquier otro procedimiento electivo. Arrow se vio pues encaminado a preguntar: ¿Surgirán solamente preferencias colectivas incoherentes cuando se aplique la regla de mayoría y procedimientos afines, o bien es éste un fenómeno intrínseco a todo procedimiento basado en votación? Para responder, Arrow pudo haber recopilado una lista de procedimientos y verificado para cada uno de ellos si ciertas configuraciones de preferencias individuales conducían a ciclos o a preferencias colectivas que presentasen algún otro rasgo inaceptable. La dificultad es que ello le hubiera obligado a considerar un número inmenso de procedimientos de agregación, distintos según el papel que le asignasen a cada votante y según el criterio con que ordenasen cada par de alternativas concretas.

Arrow tuvo por necesidad que darle al problema un enfoque axiomático. Problema que planteó como el de elegir una "constitución", es decir, una regla que asigne una ordenación de preferencias colectiva a cada una de las posibles configuraciones de preferencias individuales. Una constitución ha de especificar si una alternativa dada, cualquiera que sea, es superior, indiferente o inferior a cada una de las otras. (Dos alternativas son indiferentes cuando la sociedad las considera igualmente atractivas.) Arrow limitó el campo de constituciones posibles imponiéndoles cinco requisitos que (aducía él) son propiedades necesarias de todo método éticamente aceptable de agregación de pareceres. Seguidamente, caracterizó la clase de constituciones que verifican la totalidad de las cinco condiciones.

El primero de los axiomas de Arrow exige que la constitución sea de *carácter universal*, es decir, sea capaz de agregar cualquiera de las posibles configuraciones de preferencias del electorado. Puesto que será imposible predecir todas las pautas de conflicto que podrán surgir a lo largo de la vida de una norma de votación, explicaba Arrow, las sociedades no deben adoptar constituciones que puedan desmoronarse cuando se presenten ciertas configuraciones de preferencia en los votantes. Arrow mantenía que las sociedades deben, por el contrario, insistir en que sus constituciones sean lo suficientemente generales para elucidar todas las posibles controversias.

El segundo axioma de Arrow, el *axioma de unanimidad*, gobierna el funcionamiento de una constitución cuando no haya el menor desacuerdo entre los votantes. El axioma especifica que en las configuraciones de preferencia donde cada uno de los individuos prefiera x a y, el orden de prelación colectiva deberá situar a x por encima de y. Si se admite la idea de que la prelación que otorgue la sociedad debe ser reflejo de las preferencias de sus miembros, es difícil impugnar el postulado de unanimidad, que resuelve los que sin duda son los problemas más simples de agregación de preferencias.

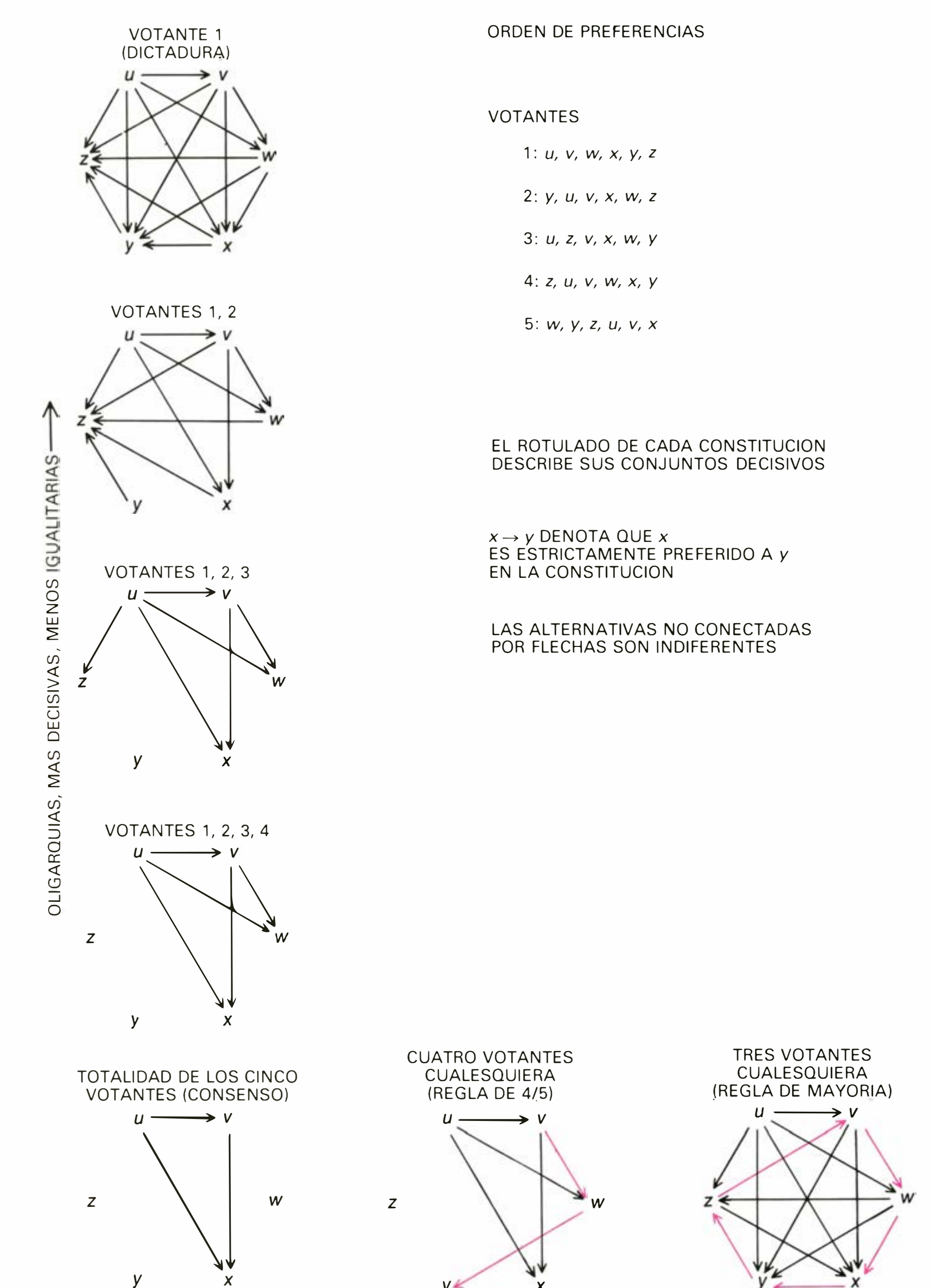
El tercero de los axiomas de Arrow, el axioma de *determinación por pares*, exige que la prelación que socialmente se confiera a cada par de alternativas ha de depender tan sólo de las prelaciones que los individuos otorguen a esas dos alternativas. Independientemente de cómo puedan variar las preferencias de los individuos respecto de otras alternativas, en tanto la prelación que cada uno de ellos otorgue a x e y permanezca invariable, también habrá de serlo la prelación colectiva de x e y. Esta condición implica, por ejemplo, que las preferencias colectivas entre Ronald Reagan y Jimmy Carter sean independientes de la prelación que los individuos den a Ted Kennedy con respecto a los dos mencionados o a Walter Mondale.

Una constitución que satisfaga la condición de determinación por pares limita la información que se requiere acerca de las preferencias de los individuos para determinar la prelación colectiva de un par de alternativas. En particular, la información sobre preferencias de los individuos acerca de opciones no disponibles no puede tener el menor efecto al determinar el orden de prelación colectiva de las sí disponibles;

propiedad que tiene importancia cuando resulta costoso o difícil educir las preferencias de prelación de los individuos. De no cumplir la exigencia de determinación por pares, la constitución tiene que especificar qué otras alternativas han de tenerse en cuenta para determinar la prelación colectiva entre x e y, así como la manera en que las preferencias de los individuos con respecto a aquellas otras han de afectar al orden de prelación colectivamente asignado a x e y.

Un procedimiento comúnmente utilizado, el procedimiento de votación por puntos, viola el postulado de determinación por pares. Supongamos, por ejemplo, tres candidaturas. Esta norma electoral asignará tres puntos al primero de los candidatos nombrado por el votante, dos al segundo y uno al tercero. La ordenación colectiva se determina entonces sumando las puntuaciones obtenidas por cada alternativa y clasificándolas por totales de sus puntos. En la comisión de tres miembros ya descrita cada candidato recibiría un total de seis puntos, con lo que la comisión sería indiferente entre ellos.

Supongamos, empero, que el orden de preferencia de Antonio cambie, pasando de x, y, z a x, z, y. Aunque ningún elector ha cambiado sus preferencias entre x e y, el método de votación por puntos declara que x es preferido a y, ya que x seguirá recibiendo un total de seis puntos, mientras que y recibirá solamente cinco. Por tanto, al votar el orden de prelación, bajo el método de voto por puntos, las preferencias colectivas entre x e y no sólo dependen de la posición relativa que les otorguen los individuos, sino también de las posiciones relativas de otras posibilidades, como z.



2. ORDENACION DE PREFERENCIAS COLECTIVAS asignadas por siete constituciones, o reglas de votación, a una misma distribución de preferencias individuales de cinco electores. Las distintas ordenaciones tipifican las distintas transacciones que se le plantean al legislador entre los requisitos de racionalidad, desigualdad de voto y resolutividad o decisividad (esto es, infrecuencia de la indiferencia colectiva). La regla de la mayoría asigna a los electores papeles simétricos, y es altamente resolutiva, pero las ordenaciones de preferencias de estos cinco votantes producen un ciclo quíntuple (*en color*) pudiendo presentarse varios ciclos más cortos. La regla de "cuatro quintos" es menos resolutiva, pero aquí no da lugar a ciclos. Sin embargo no es *P*-transitiva (coherencia de la ordenación de preferencias estrictas) pues (en notación abreviada) *vPw* (la alternativa *v* se prefiere estrictamente a la *w*) y *wPy*, mientras que *vIy* (*v* es considerada indiferente a *y*, es decir, la sociedad las considera atractivas por igual). Las oligarquías son siempre *P*-transitivas, aunque conforme se agrandan y se hacen más igualitarias son menos decisivas.

Para presentar los axiomas cuarto y quinto de Arrow conviene utilizar una notación abreviada. Se denota por P una preferencia colectiva estricta (análoga a la relación "mayor que" entre números reales), I denota la indiferencia colectiva (análoga de la igualdad entre números), y R expresa una preferencia colectiva en sentido lato (análoga al "mayor o igual" entre números). Por tanto, la expresión xRy significa que "colectivamente considerado, x es al menos tan bueno como y", es decir, que o bien xPy o bien xIy.

El cuarto axioma de Arrow es el de *completitud*: para cada par de alternativas x e y ha de ser cierto que xRy o que yRx, o ambas (en cuyo caso xIy, x e y son indiferentes). Este axioma obliga al procedimiento de agregación a ordenar cada uno de los pares de alternativas. En tanto la constitución tenga la salida de declarar indiferentes a un par cualquiera de alternativas parece que el postulado de completitud será relativamente inocuo.

El quinto axioma, que postula que R ha de ser transitiva, llamado de R-transitividad, significa, expresado simbólicamente, que si se verifica xRy, y también yRz, entonces ha de verificarse xRz. Entre las relaciones transitivas entre pares de números reales tenemos "mayor que" ($>$), "igual a" ($=$), y "mayor o igual que" ($\geqslant$). Por ejemplo, si un número x es mayor que y, y si sabemos que y es mayor que z, entonces ciertamente es x mayor que z. En el análisis económico, la completitud y la

R-transitividad son hipótesis convencionales; de los individuos que las verifiquen se dice que son "racionales". Arrow utilizó, por extensión, el término "colectivamente racionales" para describir las constituciones que verifiquen tanto el axioma de completitutd como el de *R*-transitividad. Arrow impuso la condición de *R*-transitividad para asegurar que la alternativa elegida fuera independiente del camino, plan o procedimiento seguido en la selección de la misma.

Una vez hubo definido y defendido la necesidad de este conjunto de cinco propiedades deseables, procedió a demostrar que las únicas constituciones capaces de satisfacerlas comparten todas un defecto tan sencillo cuan sorprendente: son constituciones dictatoriales. Un dictador es una persona con potestad de imponer a una sociedad su estricta preferencia sobre cualquier par de alternativas. Arrow enunció su teorema en forma ligeramente distinta. Añadió un sexto axioma, a saber, el carácter no-dictatorial de la constitución, y demostró que no existe constitución alguna que pueda satisfacer los seis axiomas. Por este motivo, suele decirse que el resultado de Arrow es un "teorema de imposibilidad".

Así pues, los redactores de normativa electoral para cuerpos legislativos, comisiones, clubs, etcétera, de considerar que estas propiedades son esenciales para toda constitución, estarán pura y simplemente faltos de suerte. Los requisitos de Arrow, en apariencia modestos, conllevan consecuencias de gran peso y deglución imposible. Tal como demuestra su teorema de imposibilidad, los cinco axiomas son extraordinariamente restrictivos; y aunque uno por uno sean muy atractivos, su combinación es perniciosa. Los teóricos han dedicado grandes esfuerzos a revisar los axiomas, buscando alguna forma de eludir la desdichada conclusión de Arrow.

Puede aducirse plausiblemente que el requisito de universalidad es demasiado ambicioso. No todas las configuraciones de preferencias posibles desde el punto de vista lógico son igualmente verosímiles. Puesto que algunas de ellas pueden ser sumamente improbables, exigirle a una constitución que agregue coherentemente toda configuración lógica y le asigne una preferencia colectiva parece una demanda innecesariamente restrictiva.

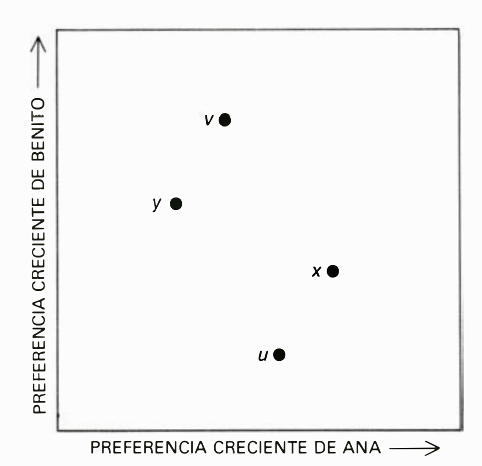
La estrategia más común para relajar este requisito ha sido tomar un determinado procedimiento, por lo común la regla de mayoría, y buscar restricciones que descarten aquellas configuraciones de preferencias que comporten preferencias colectivas intransitivas. Por ejemplo, si sólo pudieran plantearse configuraciones de preferencia triviales, donde no existen desacuerdos entre los individuos, el axioma de unanimidad determinaría las preferencias colectivas, y el problema de intransitividad no podría plantearse. De las restricciones no triviales, la más conocida es la de preferencias unimodales (*single-peaked preferences*), descubierta en el decenio de 1940 por el economista británico Duncan Black.

Se dan casos de distribuciones unimodales de preferencias cuando todos los individuos evalúan para sí cada una de las alternativas de acuerdo con un solo criterio, y en cada una de las elecciones por pares votan al candidato o alternativa más próximo a la situación considerada óptima por el individuo. Por ejemplo, cada elector podría clasificar los candidatos por afinidad con su propia concepción política, situándolos en una línea que cubra desde la extrema izquierda hasta la extrema derecha. Por tanto, si la candidatura x es más liberal que la y, y si la y lo es más que la z, una sociedad donde la distribución de preferencias fuese unimodal podría contener individuos liberales, con preferencias (x, y, z), individuos conservadores (z, y, x) e individuos moderados $(y, x, z$ o $y, z, x)$, pero no personas que valoren la opción intermedia como inferior a las extremas $(x, z, y$ o $z, x, y)$. Si fuera posible confiar en que las distribuciones de preferencia fuesen unimodales en la práctica, sería casi forzoso inclinarse por la regla de la mayoría. Por lo común, los votantes clasifican, sin embargo, las opciones electorales atendiendo a más de un criterio y la distribución de preferencias no resultará unimodal.

Con mayor generalidad, la estrategia de imponer restricciones a la configuración de preferencias sólo podrá rendir algún fruto si tales restricciones son plausibles desde la perspectiva de una teoría de la formación y estructura de las preferencias. Las ciencias sociales no han sido hasta ahora capaces de construir modelos formales del papel desempeñado por la socialización en el desarrollo de gustos o valores, ni del grado de semejanzas de preferencias que la estabilidad social exige. Así pues, a pesar del gran esfuerzo realizado, no se ha formulado ninguna caracterización de las posibles distribuciones de preferencias de los votantes que sea a un tiempo lo bastante amplia para acomodar las verdaderas preferencias de éstos y lo suficientemente estricta

para eludir la conclusión de dictadura.

La posibilidad de descartar el axioma de unanimidad no ha suscitado grandes entusiasmos. Al reconsiderar la cuestión, parece que la carga impuesta a los mecanismos de agregación de preferencias por el axioma de unanimidd es leve. Además, Robert Wilson, de la Universidad de Stanford, ha demostrado que las únicas constituciones que verifican todos los axiomas de Arrow con la excepción de la unanimidad son menos atractivas todavía que las dictaduras. La primera alternativa es la indiferencia universal, regla que hace perpetuamente indiferentes cualesquiera dos alternativas, con independencia de la consideración que les merezcan a los individuos. La segunda es la dictadura inversa, en la cual la preferencia particular de un individuo cualquier se torna al revés, y esta preferencia negativa es tomada como expresión de la voluntad colectiva. El método tan sólo podría servir a una sociedad organizada si pudiera hallarse un votante de juicios infaliblemente erróneos.



3. CONSTITUCIONES *P*-TRANSITIVAS, reglas de votación donde si *xPy* e *yPz*, entonces *xPz*. Todas las constituciones así que satisfagan además los restantes axiomas de Arrow son neutrales, esto es, la precedencia entre distintos pares de alternativas se establece por iguales criterios. "Neutralidad" significa que si una configuración concreta de preferencias de los electores implica que *u* es preferible a *v*, la misma configuración de preferencias con respecto a *x* y a *y* tiene que implicar que *x* sea preferible a *y*. Vemos en la ilustración una demostración para el caso de los electores. Supongamos que, para las alternativas *u* y *v*, la regla electoral prime a Ana sobre Benito cuando ella prefiera *u* a *v*. Con respecto a *x* e *y* hay desacuerdo. Para esta configuración es *xPu* por el axioma de transitividad; *uPv* por la hipótesis de que Ana prevalece cuando se enfrentan *u* y *v*; finalmente, *vPy* por unanimidad. De la *P*-transitividad se deduce *xPy* para esta configuración concreta de preferencias. La condición de determinación por pares, postulada por Arrow, permite una conclusión más general, a saber, que Ana prevalece cuando ella favorezca a *x* frente a *y*, independientemente de las posiciones que su preferencia, o en la de Benito, ocupen *u* y *v*. (La determinación por pares enuncia que la prelación colectiva de cualquier par de alternativas depende tan sólo de las preferencias de los electores con respecto a esas dos opciones.)

El postulado de determinación por partes recibió duros ataques durante el decenio siguiente al de la publicación del trabajo de Arrow, pero las críticas a este axioma han ido remitiendo. La primitiva defensa que Arrow hizo de su postulado se fundaba en la posibilidad de concretar constituciones que lo cumpliesen sin el lastre de tener que recoger grandes cantidades de información sobre preferencias: para clasificar las alternativas *x* e *y* no es nunca necesario averiguar qué posición ocupa *z* en la distribución de preferencias individuales. Las constituciones que violan esta regla son por lo común fastidiosas, tanto más cuantas más alternativas haya, porque, para establecer el orden de precedencia de unas cuantas alternativas viables, hay que recoger una gran masa de información. Además, puesto que las posiciones colectivamente asignadas a *x* e *y* dependen, en las constituciones que violen este axioma, de las valoraciones que los individuos hagan de otras alternativas, los votantes pueden influir en el resultado sobre *x* e *y* falseando sus preferencias sobre otras alternativas.

Probablemente sea la *R*-transitividad la menos defendible de las condiciones de Arrow. Para eludir la conclusión dictatorial de Arrow, los especialistas en teoría de votación han examinado las consecuencias en varias condiciones menos restrictivas. No es difícil demostrar que la *R*-transitividad es conjunción de dos condiciones más débiles, la *P*-transitividad (o transitividad de las preferencias en sentido estricto) y la *I*-transitividad (transitividad de la indiferencia colectiva). Por consiguiente, un procedimiento obvio para debiliar el requisito de racionalidad de Arrow es retener una de estas dos condiciones y abandonar la otra.

La transitividad de la indiferencia (*I*-transitividad) es particularmente vulnerable a la crítica, porque estudios realizados por psicólogos muestran que, en condiciones experimentales, es corriente que los individuos expresen indiferencias intransitivas. Por ejemplo, individuos que manifestan indiferencia entre *x* e *y*, y también entre *y* y *z*, pueden muchas veces preferir *x* a *z* (o al revés). Así pues, de establecerse un paralelismo entre preferencias individuales y colectivas, hay poco fundamento para exigir que las valoraciones colectivas sean *I*-transitivas.

El filósofo y economista Amartya K. Sen, de la Universidad de Oxford, ha hecho ver que, de abandonarse la *I*-transitividad y retener la *P*-transitividad, sí hay vía de escape al teorema dictatorial de Arrow, y presentó un ejemplo de constitución no dictatorial que satisface los cuatro primeros axiomas de Arrow y la *P*-transitividad. Su método, que pudiéramos llamar "regla del consenso", declara que *xPy* si y solamente si cada individuo considera que *x* es al menos tan bueno como *y* y, además, por lo menos un individuo prefiere estrictamente *x* a *y*. Por consiguiente, cuando dos individuos estén en desacuerdo sobre *x* e *y* el resultado debe ser *xIy*. Cada individuo dispone así de un "veto" que le permite bloquear una preferencia colectiva estricta opuesta a la suya. Para ver que la indiferencia colectiva no tiene por qué ser transitiva, imaginemos una comisión de dos miembros, uno con preferencia *y, x, z*, y el otro con preferencia *x, z, y*. La regla de consenso daría *xIy* e *yIz*, pero *xPz*.

El fenómeno de la intransitividad de la indiferencia puede ser reflejo de una incapacidad de los individuos para discriminar entre alternativas muy cercanas. Por ejemplo, un individuo que calibre el conservadurismo político de dos candidatos, *x* e *y*, puede encontrarlos igualmente conservadores por estar sus posiciones políticas indistinguiblemente próximas, y otro tanto puede sucederle con respecto a *y* y a *z*. Empero, los candidatos *x* y *z* pueden estar ya lo suficientemente distantes para que el individuo perciba que *x* es más conservador que *z*, y discriminar entre uno y otro.

R. Duncan Luce, de la Universidad de Harvard, ha propuesto una noción de consistencia llamada *semiorden*, a fin de plasmar en un modelo situaciones que conlleven umbrales de percepción. Las preferencias semiordenadas manifiestan *P*-transitividad, pero consienten intransitividades en la indiferencia. Imaginando que el proceso de selección colectiva sea un proceso de agregación de los juicios de valor de los individuos a fin de formar juicios de valor colectivos, la noción de discriminación imperfecta puede aplicarse tanto a preferencias colectivas como a las individuales.

El semiorden es, sin embargo, un requisito de racionalidad más estricto que la mera *P*-transitividad, hecho que tiene importantes consecuencias para los procedimientos de agregación de preferencias. Como ya hemos puesto de manifiesto en otro lugar, al exigir que una constitución produzca preferencias colectivas semiordenadas y verifique los otros axiomas de Arrow, se sigue implicando que será dictatorial. Por consiguiente, la justificación que para debili-

tar la *R*-transitividad nos proporciona la teoría del umbral de discriminación, aunque atractiva y plausible, no consigue eludir la penosa conclusión de Arrow.

La principal justificación que K. J. Arrow dio para su axioma de *R*-transitividad fue que la elección efectuada a partir de algún conjunto de alternativas debería ser independiente de la agenda u orden de procedimiento mediante el cual se efectúa la votación. Vale la pena notar que el requisito de independencia respecto al procedimiento nos permite argumentar directamente que la forma de racionalidad exigible de las preferencias colectivas sea la *P*-transitividad. Charles R. Plott, del Instituto de Tecnología de California, ha propuesto una definición formal de independencia del camino, y ha demostrado que las constituciones que verifiquen su definición habrán todas de ser *P*-transitivas. Además, toda constitución *P*-transitiva satisface dicha condición. Así pues, aunque la primitiva condición de racionalidad exigida por Arrow garantiza que la elección final no depende del camino seguido, otra condición menos restrictiva permitiría lograr el mismo objetivo.

La regla del consenso de A. M. Sen demuestra que existen constituciones *P*-transitivas no dictatoriales. Mas, como el propio Sen reconoció, la regla del consenso no es solución atractiva del problema de Arrow; con demasiada frecuencia no es decisoria. Cada vez que un par de individuos manifiesten preferencias estrictas contradictorias ante un par de alternativas –situación ubicua donde haya intereses en conflicto– la regla del consenso declarará indiferentes ambas alternativas. Una constitución que establezca indiferencia entre las preferencias colectivas cada vez que las de dos individuos entren resulta conflicto es virtualmente inútil.

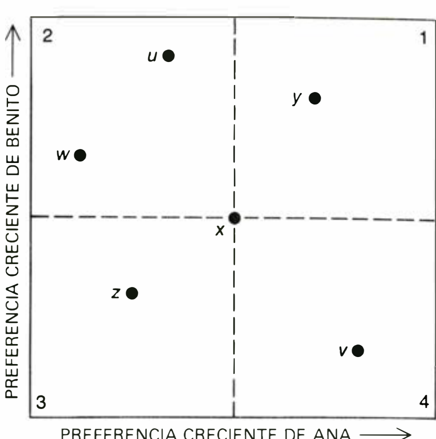
¿Podrán descubrirse otras constituciones *P*-transitivas más atractivas? El filósofo Allan Gibbard, de la Universidad de Michigan, ha probado que no. Gibbard demostró que en todo proceso *P*-transitivo que obedezca a los restantes axiomas de Arrow existe un conjunto privilegiado de individuos, que Gibbard llama "oligarquía". Esta oligarquía tiene, como grupo, capacidad para imponer su preferencia unánime y estricta sobre cualquier par de alternativas. Además, cada miembro de la oligarquía tiene, individualmente, el poder de anular con su veto cualquier preferencia colectiva estricta contraria a la suya. Siempre que uno de los oligarcas prefiera estrictamente *x* a *y* será imposible que *yPx*. Así pues, las dictaduras son oligarquías de un solo individuo, mientras que la regla del consenso equivale a una oligarquía formada por la sociedad entera.

No todas las constituciones oligárquicas distribuyen desigualmente el poder, como pone de manifiesto la regla de consenso. Conforme más y más individuos van instalándose en la oligarquía, más equitativa va haciéndose la distribución del poder. Empero, cuanto mayor es la oligarquía mayor es la probabilidad de que se produzcan situaciones de indecisión, ya que los conflictos entre oligarcas implican la indiferencia colectiva. Puesto que la independencia del procedimiento exige la *P*-transitividad, el teorema de Gibbard conlleva que esta atractiva propiedad tan sólo puede adquirirse al precio de hacer el sistema poco irresolutivo, o de hacer desigual la fuerza del voto de los electores. Para escapar al dilema de tener que optar entre las desigualdades inherentes a una pequeña oligarquía y la inoperancia de las grandes será preciso debilitar todavía más la noción de racionalidad colectiva, y no exigir siquiera que sea *P*-transitiva.

Las constituciones *P*-transitivas que satisfacen los restantes axiomas de Arrow tienen otro defecto más, el de ser inflexibles, en el sentido de que, excepto cuando todos los miembros de la oligarquía manifiesten su indiferencia, no pueden imponer requisitos sobre las preferencias colectivas estrictas que sean distintos para diferentes pares de alternativas. Por el contrario, es forzoso que traten a todos los pares de alternativas de modo simétrico o neutro. Así, si Antonio (pero no Benito ni César) se ve especialmente afectado por un cambio de la política *y* a la política *x*, puede ser apropiado darle a Antonio el derecho de vetar la ordenación *xPy*. Puede ser indeseable, empero, que disponga de tal poder sobre otros pares. Las constituciones *P*-transitivas que verifiquen los restantes axiomas de Arrow no pueden, sin embargo, gozar de tanta flexibilidad.

Bajo las consituciones oligárquicas *P*-transitivas que satisfacen los restantes axiomas de Arrow, los grupos (o individuos) con poder para imponer su voluntad sobre un par de alternativas habrán de tener igual poder para cualesquiera otros pares. Análogamente, un individuo o grupo con poder de veto sobre un par determinado de alternativas habrá de poderlo ejercer igualmente sobre otro par cualquiera. Si una determinada configuración de preferencias entre *x* e *y* implica que *x* es colectivamente preferible a *y* (el postulado de determinación por pares garantiza que ninguna información adicional sobre preferencias de los individuos puede pesar sobre la valoración colectiva que se haga de *x* e *y*), la misma configuración de preferencias entre *z* y *w* implicará que *z* sea colectivamente preferido a *w*. Para tratar asimétricamente, o de formas no neutrales, distintos pares de alternativas será preciso que las constituciones verifiquen requisitos menos estrictos que el impuesto por la *P*-transitividad.

Entre las condiciones menos restrictivas que la *P*-transitividad tenemos la de *aciclicidad*, esto es, la no existencia de ciclos de preferencia colectiva de ninguna longitud. Por ejemplo, con tres alternativas, la prelación colectiva *xPy*, *yPz*, *xIz* es acíclica (pues no existen en ella ciclos de preferencia estric-

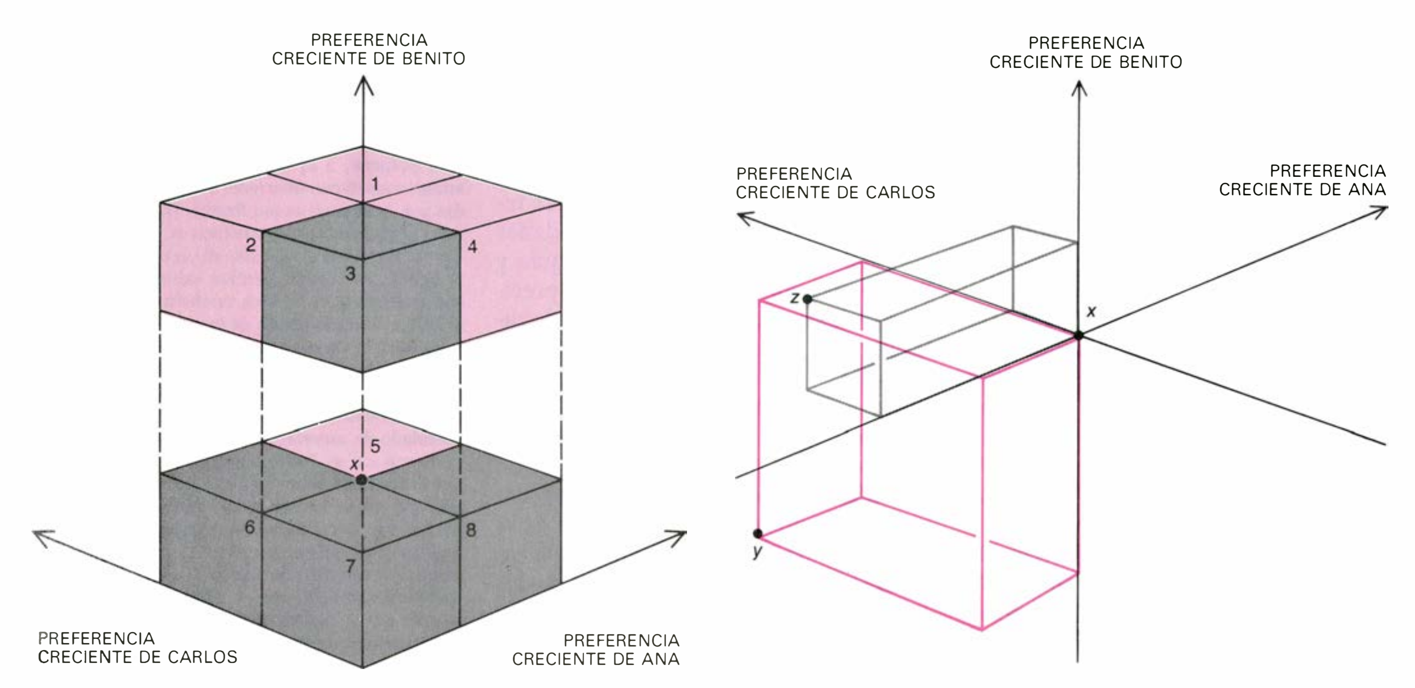


4. TEOREMA DE DICTADURA, de Arrow, demostrado aquí para dos votantes. Tenemos representada una posible configuración de preferencias de Ana y Benito; el cuadrante en que aparece cada alternativa depende de cómo la sitúen los electores con respecto a la alternativa *x*. El cuadrante *1* contiene las alternativas (como la *y*) preferidas a *x* por los dos votantes; el *2*, las que Benito antepone, y Ana pospone, a *x*; el cuadrante *3* recibe las que ambos consideran inferiores a *x*, y el *4*, las preferidas por Ana, pero no por Benito. De la neutralidad de las constituciones transitivas se deduce que todas las alternativas situadas en un mismo cuadrante habrán de recibir idéntica valoración colectiva con respecto a *x*. Implica también que el razonamiento es independiente de la alternativa concreta utilizada para la descomposición en cuadrantes. El cuadrante *2* no puede ser indiferente con respecto a *x*, pues entonces, al mismo tiempo *uIx* y *xIw*, y por la transitividad de *I*, *uIw*, lo que contradice el postulado de unanimidad, ya que ambos votantes prefieren *u* a *w*. Análogamente, el cuadrante *4* no puede ser indiferente respecto de *x*. Los cuadrantes *2* y *4* no pueden ser ambos preferidos, ni ambos preteridos, con relación a *x*, pues la neutralidad implica que si *uPx* entonces *xPz*, ya que las valoraciones que los votantes hacen de *u* y *x*, y de *x* y *z*, son iguales. Puesto que el axioma de unanimidad implica que el cuadrante *1* sea preferido a *x*, y el *3* sea indiferente, sólo hay dos constituciones coherentes con los axiomas de Arrow: si *4* es preferido a *x* y *2* inferior, Ana será una dictadora; mientras que si el preferido es *2* y el relegado el *4*, será Benito el dictador. La demostración se debe a Charles Blackorby, David Donaldson y John Weymark.

ta), pero viola la *P*-transitividad (pues las dos primeras prelaciones exigirían que *xPz*). Entre las preferencias colectivas descartadas por el requisito de acircularidad tenemos el ciclo de tres alternativas explicado en la paradoja electoral.

La propiedad de aciclicidad es atractiva, sobre todo cuando el procedimiento adoptado para elegir alternativas partiendo de ordenaciones colectivas entre pares consiste en ir seleccionando aquellas alternativas que no serían aisladamente derrotadas por ninguna otra. La aciclicidad garantiza que en todo conjunto factible existe siempre al menos una de tales alternativas, y es la menos restrictiva de las condiciones suficientes para ello. En el ejemplo acíclico anterior, pongamos por caso, la alternativa *x* no es derrotada por ninguna de sus rivales, mientras que con la preferencia colectiva cíclica *xPy, yPz, zPx* cada alternativa es derrotada por alguna otra. Sin la propiedad de aciclicidad, la selección colectiva apropiada dista de estar clara.

Las constituciones acíclicas pueden satisfacer los restantes axiomas de Arrow sin necesidad de ser neutrales, pudiendo así tratar en forma asimétrica las distintas alternativas. Los procedimientos asimétricos son de uso frecuente y bien admitido. Por poner un ejemplo (en ocasiones cíclico), en el Senado norteamericano basta la mayoría simple para aprobar una disposición ordinaria, pero las mociones para limitar el tiempo de debate exigen ya mayoría de tres quintos, y las propuestas de enmienda a la Constución requieren dos tercios. Que la neutralidad sea o no requisito deseable en una constitución depende de la naturaleza de las alternativas. Cuando se trata de seleccionar entre aspirantes a un cargo, la condición de imparcialidad parece más



5. CONSTITUCIONES NEUTRALES, que pueden representarse por ocho octantes, de haber solamente tres electores. La propiedad de neutralidad, que suponemos para las constituciones acíclicas y que es necesariamente cierta si la constitución es *P*-transitiva, asegura que todas las alterantivas encasilladas en un mismo octante recibirán idéntica consideración colectiva (superior, indiferente o inferior) con respecto a la *x*. Si un octante, el *2*, por ejemplo, es colectivamente preferido a *x*, un opuesto (*8*) tiene que ser inferior; y si un octante es colectivamente indiferente a *x*, su opuesto ha de ser indiferente también. Puesto que el axioma de unanimidad obliga a la constitución a preferir el octante *1* a la alternativa *x*, y la *x* frente al octante *7*, todas las constituciones neutrales posibles quedan totalmente descritas en cuanto se especifique la valoración colectiva asignada a los octantes *2, 3* y *4* con respecto a *x*. Puesto que hay, para cada uno, tres posibilidades, en total son 27 las constituciones neutrales posibles con tres individuos. Para constituciones *P*-transitivas que satisfagan los restantes axiomas de Arrow puede demostrarse que, si *x* recibe más apoyo frente a *y*, la posición de *x* no puede resultar perjudicada frente a *y* en la prelación colectiva. Por ejemplo, supongamos que Ana cambie su preferencia, llevando a *z* desde el octante *3* al *4*, es decir, en lugar de anteponer *x* a *z*, ella antepone ahora *z* a *x*. Entonces la valoración colectiva de *4* con respecto a *x* debe ser al menos tan favorable como la de *3*; si *3* es indiferente, *4* debe ser indiferente o preferido. Ciertas constituciones no *P*-transitivas infringen esta asociación positiva entre valoración individual y valoración colectiva. Puesto que tales constituciones no resultan atrayentes, suele exigirse el requisito de asociación positiva para las constituciones no *P*-transitivas. Puesto que desplazar una alternativa desde *3* hasta *2* o *4* es reflejo de una mejora en la valoración que algún votante hace de ella al compararla con *x*, la asociación positiva implica que si *3* es preferido a *x* también lo han de ser *2* y *4*. Si *3* es indiferente a *x, 2* y *4* deben ser indiferentes o preferidos a *x*. Si *2* es inferior, *8* ha de ser preferido y, por tanto, *4* debe ser preferido, por asociación positiva. Razonamientos así reducen el número de constituciones tripersonales que verifican los requisitos de neutralidad, asociación positiva y los cuatro primeros axiomas de Arrow, de 27 a 11. Se muestra la regla de mayoría, con los octantes preferidos a *x* (*color*) y los preteridos (*gris*); ninguno es indiferente.

6. LAS CONSTITUCIONES *P*-TRANSITIVAS y las constituciones acíclicas pueden caracterizarse mediante la configuración de preferencias individuales aquí mostradas. Hay 11 constituciones tripersonales que verifiquen los requisitos de neutralidad, asociación positiva y los cuatro primeros axiomas de Arrow; las vemos representadas en las ilustraciones 5 y 7. En cada constitución ha de existir un individuo con facultad de veto. Para demostrarlo bastará ver que con la regla de mayoría pueden suscitarse ciclos. En la regla de mayoría, los octantes *1, 2, 4* y *5* son colectivamente preferidos a *x*, mientras los *3, 6, 7* y *8* son inferiores. Para la configuración de preferencias individuales aquí mostradas, la regla de la mayoría genera un ciclo: *xPy* (porque *y* está en el octante *6*), *yPz* (pues *y* tiene respecto de *z* la misma relación que el octante *5* respecto de *x*) y *zPx* (porque *z* está en el octante *2*). Por tanto, la constitución no podrá ser acíclica sin otorgar a ciertos individuos la facultad de veto. Alland Gibbard, de la Universidad de Michigan, ha demostrado que las constituciones *P*-transitivas que satisfagan los cuatro primeros axiomas de Arrow son oligárquicas, esto es, dan poder a un único grupo de votantes para imponer a la sociedad su preferencia conjunta, unánime y estricta, y confieren además a cada miembro de ese grupo, individualmente, la facultad de vetar preferencias colectivas estrictas contrarias a la suya. Las siete constituciones inferiores de la figura 7 son oligárquicas. Para demostrar el resultado de Gibbard en el caso de los tres votantes sólo se necesita ya mostrar que las tres constituciones de lo alto infringen la *P*-transitividad. Puesto que sólo difieren entre sí en los nombres de los votantes, se razonará solamente en un caso, el segundo. Como *z* es preferida en el octante *2, zPx*. La alternativa *y* se encuentra en el octante *6*, opuesto al octante preferido *4*, así que *xPy*. La *P*-transitividad exigiría que *zPy*. Mas *z* tiene respecto de *y* la misma relación que el octante *3* (ahora indiferente) respecto de *x*, contradiciendo la neutralidad de las constituciones *P*-transitivas y demostrando el resultado. El requisito de *P*-transitividad conlleva pues oligarquías. Reforzando el requisito de racionalidad a la *R*-transitividad, que pide que la preferencia débil sea transitiva, quedan suprimidas las oligarquías de más de un miembro, dejando tan sólo las tres dictaduras. Las cajas de este diagrama sirven para localizar a *y* y *z* en los octantes apropiados de la ilustración 5. (Los dibujos son de Ilil Arbel.)

atrayente que otras que pudieran primar a determinados candidatos. Pero cuando las consecuencias de las distintas alternativas sean asimétricas, puede resultar ventajosa cierta parcialidad de las reglas que tienda a dificultar las de más drásticos resultados. El sistema de enjuiciamiento criminal norteamericano nos muestra ejemplos en ambos extremos del espectro: cuando el caso es de poca monta no suele recurrirse a juicio público ante jurado; las sentencias de muerte son automáticamente revisadas por tribunales de apelación.

Al relajar la exigencia de racionalidad desde la *P*-transitividad a la mera aciclicidad se hacen posibles muchas constituciones desprovistas de oligarquías. Todas las nuevas constituciones adolecen, sin embargo, de uno de los peores inconvenientes de las reglas oligárquicas. Todas las constituciones acíclicas que verifiquen los restantes axiomas de Arrow otorgan a un cierto grupo o individuo amplias facultades de veto.

Típica entre estas constituciones acíclicas era la regla de votación del Consejo de Seguridad de las Naciones Unidas hasta 1965. El Consejo de Seguridad estaba entonces formado por cinco miembros permanentes y seis no permanentes. Para que una moción prosperase era necesario que recibiese al menos siete votos afirmativos y ningún voto negativo de ninguno de los miembros permanentes. Así pues, cada miembro permanente del Consejo de Seguridad disponía de veto; ninguna moción podría prosperar si uno tan sólo de los cinco miembros permanentes se oponía a ella. Estas cinco naciones no constituían, sin embargo, una oligarquía, pues hacía falta el apoyo adicional de miembros no permanentes para lograr una preferencia colectiva estricta.

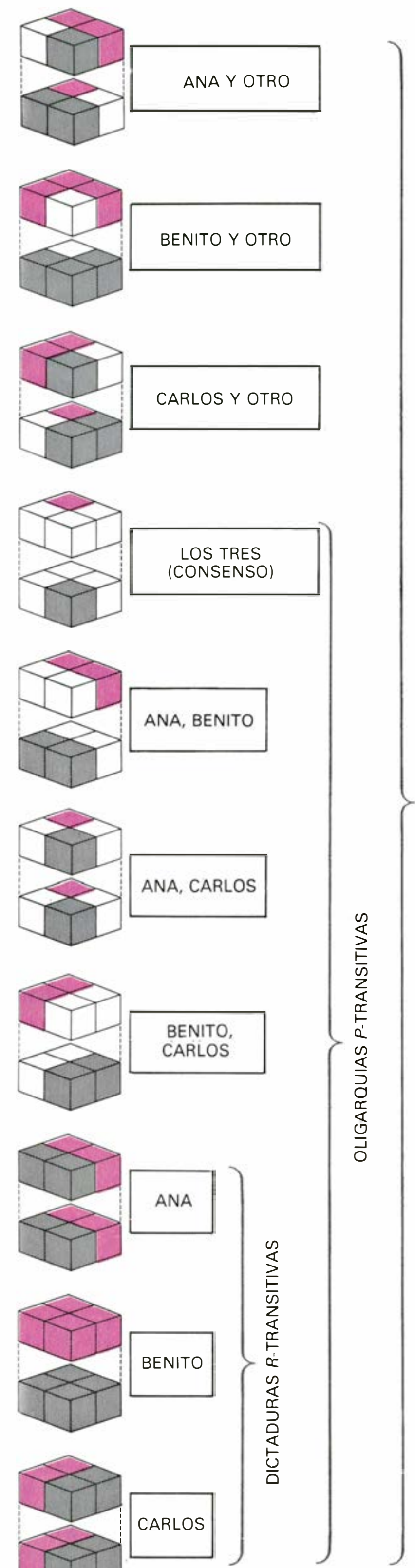
Bajo las constituciones acíclicas un individuo o grupo puede tener facultades para vetar ciertas alternativas, pero no otras. La amplitud de la facultad de veto debe, por consiguiente, quedar descrita en términos del número de pares sobre los que algún individuo o grupo puede ejercer su veto. Como los autores han demostrado recientemente, cuando el número de alternativas es grande respecto del número de electores, al menos uno de ellos debe quedar facultado para vetar gran número de alternativas. Con mayor precisión, conforme la relación entre número de pares de alternativas y número de votantes crece ilimitadamente, la proporción del número de pares de alternativas sobre las que algún individuo tendrá poder de veto va acercándose a la unidad.

Incluso cuando haya menos alternativas que individuos, su relación numérica es crítica. La *R*-transitividad y la *P*-transitividad son condiciones de racionalidad aplicables a ternas de alternativas. La aciclicidad, por otra parte, descarta la posibilidd de ciclos, sea cual fuere su longitud, y la inexistencia de ciclos de longitud tres no descarta que puedan existir otros de longitud cuatro o mayor. Conforme aumenta el número de alternativas van haciéndose posibles ciclos más y más largos.

Con dos alternativas solamente, la regla de mayoría no puede engendrar ciclos. Un ciclo exigiría que más de la mitad de los votantes prefiriesen x a y, y que, al mismo tiempo, más de la mitad prefiriesen y a x; por consiguiente, al menos un individuo tendría simultáneamente que preferir x a y e y a x, lo cual es claramente imposible. Cuando hay tres o más alternativas, como nos muestra la paradoja de Condorcet, ningún individuo tiene necesidad de estar simultáneamente conforme con todos los eslabones del ciclo de preferencias colectivas, y así el ciclo no contradice la transitividad de la ordenación de preferencias de ninguno de los individuos.

Un comité de cinco miembros nos permite ilustrar mejor el crítico papel de la proporción entre alternativas e individuos. Consideremos una constitución que requiera cuatro votos afirmativos para otorgar la preferencia estricta –una constitución intermedia entre la regla de la mayoría y la regla de consenso. ¿Podría suscitarse el ciclo xPy, yPz, zPx bajo esta constitución de "mayoría por cuatro quintos"? No, pues las preferencias de los individuos son transitivas, y al menos cuatro votantes han de estar de acuerdo en cada uno de los tres eslabones del ciclo de preferencia colectiva, lo cual es imposible.

Con un comité de cinco miembros y tres alternativas, la regla de mayoría por cuatro quintos presenta ventajas sobre la mayoría ordinaria o la regla del consenso. A diferencia de la primera, no pueden presentarse ciclos; a diferencia de la segunda, ningún miembro dispone de veto. No obstante, tales ventajas tienen un precio. La regla de los cuatro quintos es menos racional que la del consenso, pues aunque las preferencias colectivas que determina son acíclicas, no son *P*-transitivas y, por tanto, no son independientes del procedimiento u orden del día que condujo a ellas. Y es menos decisoria que la regla



7. CONSTITUCIONES NEUTRALES, de tres votantes, que satisfacen diversas condiciones de racionalidad. Cada una de estas figuras de ocho octantes ilustra la valoración de alternativas con respecto a una alternativa fija, *x*, según constituciones concretas. Las alternativas de los octantes coloreados son superiores a *x*; las de octantes en gris, inferiores, y las de octantes en blanco, indiferentes a *x*. Los números de las cajas señalan qué votante o grupo es decisivo en cada situación particular.

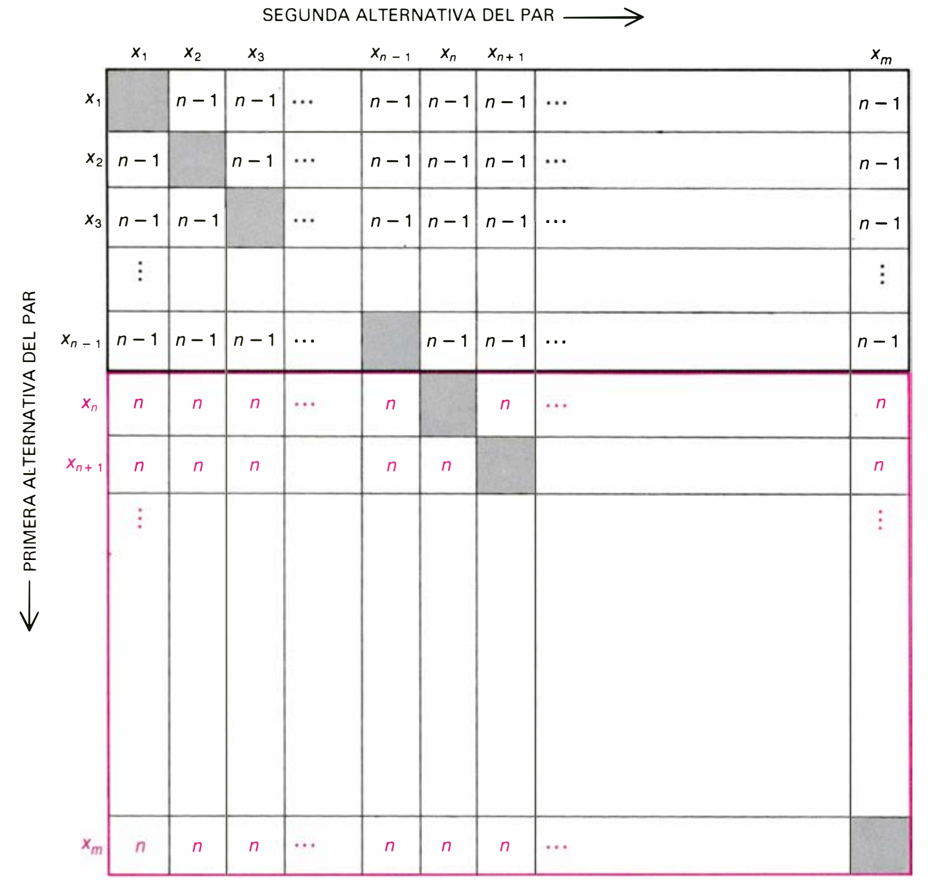
de mayoría ordinaria: aunque ningún individuo dispone de veto, cualquier grupo de dos miembros puede evitar preferencias colectivas contrarias a las suyas propias.

En el caso de constituciones acíclicas, el tamaño del grupo mínimo que dispone de facultad de veto depende de la proporción relativa de alternativas e individuos. Cuando sólo hay dos alternativas, como claramente muestra la regla de mayoría, el mínimo grupo con facultad de veto es la mitad del electorado. Cuando hay tantas alternativas cuantos votantes, bajo toda regla acíclica tiene que existir al menos un individuo que por sí solo disponga de poder para vetar ciertos pares de alternativas. En casos intermedios, como pone de manifiesto la constitución que estipula la mayoría de cuatro quintos, algunos grupos minoritarios tienen facultad de veto. Las constituciones que confieren poder de veto a gran número de pequeños grupos conducirán con frecuencia a puntos muertos, más que a decisiones.

Los igualitaristas más estrictos aducirían ahora que, si algunos individuos han de disponer de poder de veto, todos los individuos han de tener esa misma facultad. Empero, cuanto mayor sea la población con derecho a veto, tanto más frecuentemente se darán casos de indiferencia colectiva, pues la indiferencia se presenta cada vez que dos individuos con facultad de veto en direcciones apropiadas valoran x e y en sentidos opuestos. El atractivo de las constituciones acíclicas será en mayor o menor grado placentera dependiendo de la relación entre alternativas y votantes. Cuando el número de alternativas sea sólo ligeramente inferior al de individuos, se tolerarán amplias facultades de veto a grupos "pequeños", si bien no es forzoso que haya de poseerlas ningún individuo por sí solo. Conforme aumenta el número de alternativas va disminuyendo el tamaño del mínimo grupo facultado para vetar que viene exigido por el carácter acíclico de la constitución. Con iguales números de alternativas y votantes, al menos un individuo deberá poder vetar ciertos pares de alternativas; y al aumentar todavía más el número de alternativas, la proporción de pares sobre los cuales el individuo tiene facultad de veto tiende a la unidad.

Los "teoremas de imposibilidad", que comenzaron con la famosa tesis de Arrow, imponen restricciones a la libertad de las sociedades para determinar reglas de adopción colectiva de decisiones. Tales restricciones son verdaderamente serias. Tres objetivos ampliamente deseados (racionalidad colectiva, decisividad e igualdad en la fuerza del voto) se encuentran en conflicto irreconciliable. Si la sociedad renuncia a ser colectivamente racional, aceptando pues como mal necesario la arbitrariedad y manipulabilidad de procedimientos irracionales, es probable que la fórmula elegida sea la de votación por mayoría, pues esta regla consigue los restantes objetivos. Si la sociedad insiste en conservar cierto grado de racionalidad colectiva, puede lograr un voto igualitario adoptando la regla de consenso, pero sólo al costo de la irresolutividad más extrema. La sociedad puede ir aumentando su decisividad concentrando el poder de veto cada vez en menos manos; pero la regla de máxima capacidad decisoria, la dictadura, es también la menos igualitaria.

Poco consuelo queda pues para quienes aspiren a diseñar procedimientos ideales de selección colectiva. Empero, toda sociedad tiene forzosamente que tomar decisiones colectivas, y diseñar procedimientos de votación, por muy imperfectos que puedan ser. El análisis axiomático, persistiendo en la línea de investigación trazada por Arrow hace 33 años, ha producido una comprensión más profunda de los métodos de votación existentes, y puede que llegue a proporcionarnos otros mejores. Demuestra también que las oportunidades de mejora están seriamente limitadas. Son inevitables los compromisos.



	x_1	x_2	x_3		x_{n-1}	x_n	x_{n+1}		x_m
x_1		$n-1$	$n-1$	…	$n-1$	$n-1$	$n-1$	…	$n-1$
x_2	$n-1$		$n-1$	…	$n-1$	$n-1$	$n-1$	…	$n-1$
x_3	$n-1$	$n-1$		…	$n-1$	$n-1$	$n-1$	…	$n-1$
	⋮								⋮
x_{n-1}	$n-1$	$n-1$	$n-1$	…		$n-1$	$n-1$	…	$n-1$
x_n	n	n	n	…	n		n	…	n
x_{n+1}	n	n	n		n	n			n
	⋮								⋮
x_m	n	n	n	…	n	n	n	…	

8. UNA AMPLIA FACULTAD DE VETO es característica de todas las constituciones acíclicas que verifiquen los cuatro primeros axiomas de Arrow. La constitución aquí representada, definida para n votantes y m alternativas, puede ser descrita según el número mínimo de votantes cuyo acuerdo es necesario para establecer una preferencia colectiva estricta entre elementos de un par ordenado de alternativas distintas. Este número aparece en la casilla correspondiente al par ordenado. Para lograr preferencia colectiva estricta en cualquiera de los $(m - n + 1)(m - 1)$ pares de la región roja de la tabla hace falta el consentimiento unánime de todos los votantes; por consiguiente, en esta región cada votante dispone de veto. Para ver que esta constitución no puede producir ciclos de preferencia colectiva, imaginemos lo contrario. Al menos uno de los votantes ha de estar disconforme con cada uno de los eslabones del ciclo de preferencia colectiva que yace en la región en negro de la tabla. Así pues, incluso siendo diferente el votante disconforme con cada eslabón, algún votante habrá de estar conforme con todos ellos, ya que a lo sumo hay $n - 1$ pares del ciclo en esa región. Ese mismo individuo habrá de estar conforme también con cada uno de los eslabones coloreados del ciclo. Por estar conforme con todos los eslabones del ciclo, las preferencias de tal individuo son cíclicas. Esta contradicción prueba que la constitución no admite ciclos. Cuando m es mayor que n esta regla le confiere a cada elector poder de veto sobre $(m - n + 1)(m - 1)$ pares de alternativas. Al hacerse grande la razón de m a n, la regla es poco atractiva porque es muy poco resolutiva. Mas, como los autores han demostrado recientemente, toda regla acíclica que verifique los cuatro primeros axiomas de Arrow otorga a algún elector facultad de vetar al menos este número de pares si m es mayor que n. Aunque la regla es poco atrayente, en estas circunstancias no hay otra mejor.

Iglesias noruegas de pies derechos

En el siglo x, los noruegos mezclaron elementos paganos y cristianos en sus iglesias de madera. Algunas siguen aún en pie, demostrando que los edificios de madera, correctamente diseñados y mantenidos, pueden ser permanentes

Petter Aune, Ronald L. Sack y Arne Selberg

Abundan las pruebas de que, en la antigüedad, numerosas estructuras se construían de madera, y no sólo de piedra, ladrillo o argamasa. Sin embargo, casi la totalidad de las estructuras antiguas, e incluso medievales, que han llegado hasta nosotros están construidas en piedra. La razón es evidente. La piedra, el ladrillo y la argamasa, aunque lejos de ser indestructibles, en igualdad de condiciones perduran más que la madera, que es vulnerable a la pudrición, a los insectos nocivos, al fuego y al abandono. Una excepción a esta norma son las iglesias noruegas de pies derechos, un tipo de construcción en madera que data de fines del siglo x. Tal excepción es muy significativa.

Los soportes de estas iglesias son pies derechos de madera. Entre los siglos xi y xiv se construyeron en Noruega numerosas iglesias de ese tipo. Se combinaron elementos de origen pagano con los de procedencia cristiana. Resultado de tal combinación es un diseño de apariencia primitiva, con un exótico exterior y un apacible interior, de proporciones básicamente humanas. Sin embargo, el aspecto más destacado de estas iglesias es la durabilidad de su estructura. ¿Cómo se ha logrado que algunos edificios de madera de este tipo hayan sobrevivido más de 800 años?

Tres ingenieros de estructuras hemos examinado más de la mitad de las iglesias de pies derechos que aún se conservan. Hemos concluido que son tres las razones que han permitido a los edificios resistir tanto tiempo. En primer lugar, el extraordinario cuidado que se tuvo en preparar la madera, con un especial método de secado que facilitó a los antiguos constructores noruegos unos materiales muy superiores a la actual madera de construcción. En segundo lugar, el diseño, en su conjunto, es excelente y muy adecuado a los lugares desapacibles en que muchas de estas iglesias se construyeron. La estructura es capaz de soportar los empujes verticales originados por la propia masa de la iglesia y por el peso de la nieve. Además, le suministra una combinación de fuerza y rigidez capaz de resistir las fuerzas laterales ejercidas por el fuerte viento. En tercer lugar, innovaciones estructurales, pequeñas aunque significativas, han servido para proteger la madera e impedir su deterioro. La conclusión es que con un diseño apropiado y poniendo especial cuidado en preparar el material constructivo, las grandes estructuras de madera pueden ser permanentes. En realidad, la historia de las iglesias de pies derechos pone de manifiesto que debe revisarse la teoría, generalmente aceptada, de que la madera es adecuada sobre todo para pequeños edificios o para construcciones de las que no se espera una larga vida.

Las iglesias de pies derechos incorporan dos tipos fundamentales de estructuras de madera: la de armadura y la de viguería. En una estructura ideal de armadura cada elemento está sometido a una fuerza de tracción o de compresión, pero nunca a ambas a la vez. Es más, no se ejerce otro tipo de cargas sobre los elementos. Un sencillo ejemplo de esta estructura es un ensamblaje triangular con el vértice del triángulo hacia arriba, tal y como se encuentra en un tejado. Una fuerza vertical hacia abajo aplicada en el vértice sitúa a los dos elementos diagonales en compresión y, al de base, en tensión. Una fuerza de estas características puede deberse tanto al peso del propio tejado como al de la nieve en él depositada.

Por otra parte, en una estructura de viguería las fuerzas de tracción y compresión pueden darse a la vez en un mismo elemento. Igualmente, pueden existir otros tipos de fuerzas en dicha estructura, tales como las ejercidas por el deslizamiento o el alabeo. En su forma más simple, una estructura de viguería está formada por un dintel que descansa sobre dos soportes verticales. Este tipo de ensamblaje puede formar parte del armazón de una casa. El peso de la viga horizontal y las cargas que soporta se manifiestan como una fuerza hacia abajo que comprime los miembros verticales. Estos, no obstante, tienden a curvarse hacia afuera, por lo que están sujetos a una fuerza de compresión y también a una fuerza de alabeo.

Las estructuras adinteladas con vigas existen desde la prehistoria. Su trazado puede establecerse con sólo disponer un elemento horizontal sobre dos soportes. El esquema de armadura tiene, igualmente, una larga historia. Se recuerda a los arquitectos romanos fundamentalmente por sus trabajos en ladrillo, argamasa y mármol, pero también realizaron impresionantes obras en madera. La Columna Trajana, en los Foros Imperiales de Roma, se erigió en el 114 d. de C. para conmemorar la victoria del emperador sobre los pueblos dacios, habitantes de lo que hoy es Rumanía. Entre los temas representados en la columna figura uno que muestra cómo en el año104 d. de C. el arquitecto imperial Apollodoro de Damasco construyó un puente de madera sobre el Danubio. El esquema del puente in-

1. IGLESIA DE PIES DERECHOS DE FANTOFT, un ejemplo bien conservado de las más antiguas iglesias de ese tipo. Construida alrededor de 1200 en Fortun, un pueblo de la cabecera del fiordo Sogne, se desmanteló y reconstruyó en 1883 en Fantoft, cerca de Bergen. La iglesia de pies derechos recibe su nombre de los pilares de madera que son sus principales soportes. En el exterior de la iglesia, los constructores combinaron motivos paganos y cristianos. Las cabezas de dragón situadas sobre los gabletes, similares a las que aparecen en las proas de las naves vikingas, destacan más que las cruces, instaladas más abajo. El edificio consta de una nave rectangular (*primer término*) y un pequeño presbiterio (*al fondo*). La iglesia de Fantoft es de diseño basilical: la nave central es más alta que las laterales. De los tres tejados principales, el mayor cubre la nave, el mediano las alas y el menor el pórtico. (Fotografía de Sack.)



corpora estructuras de armadura. El modo en que se han combinado las estructuras adinteladas de viguería y las de armadura en las iglesias de pies derechos es especialmente ilustrativo de por qué estos edificios son tan resistentes.

La madera continuó empleándose en Europa como material de construcción tras la caída del Imperio Romano. Los fabulosos barcos construidos por los vikingos demuestran que las técnicas de construcción en madera se desarrollaron enormemente en los países nórdicos, durante la Alta Edad Media. Además, los barcos vikingos mantuvieron a los pueblos nórdicos en contacto con los centros culturales, permitiéndoles conocer las técnicas que sobre la construcción de edificios se habían desarrollado en otras partes de Europa.

La influencia de tal tecnología se fortaleció en Noruega gracias a la introducción del cristianismo. La religión cristiana llegó a Noruega más tarde que a la mayoría de los otros países europeos. El primer rey que intentó cristianizar Noruega fue Haakon el Bueno, quien reinó desde 935 hasta 961. Haakon construyó tres iglesias en el distrito de Møre; poco después, las iglesias fueron incendiadas, y asesinados los sacerdotes que había traído de Inglaterra. Los esfuerzos misioneros de varios reyes posteriores sufrieron también un feroz rechazo. Hasta el reinado de Olav Kyrre, quien gobernó desde 1066 hasta 1093, no se estableció de forma duradera una iglesia oficial en Noruega. Olav fijó sedes episcopales en Trondheim, Bergen y Oslo, y probablemente inició un amplio programa de construcción de iglesias.



2. **IGLESIA DE GOL, construida entre 1200 y 1250 y trasladada, en 1884, al Museo noruego de Arte Popular de Bygdø, cerca de Oslo. La fotografía está tomada desde la nave hacia el presbiterio, donde se encuentra el altar. La iglesia de Gol también presenta esquema basilical. La nave central está limitada por los pies derechos interiores, 14 soportes de madera de ocho metros de altura. Seis de ellos no se apoyan en el suelo. Los elementos curvos situados bajo los pies derechos acortados permiten que los empujes se transmitan a los soportes de altura completa de los ángulos, los cuales transfieren las cargas a la infraestructura. El conjunto de los pies derechos logra su estabilidad mediante una serie de horquillas, la arquería (situada aquí por encima del Crucificado), y gracias a unos tirantes entrelazados llamados cruces de San Andrés (detrás del Crucificado). Los pies situados al fondo configuran el presbiterio.**

De este modo, durante los siglos x y xi, el número de iglesias se incrementó en Noruega, aunque no sin ciertas dificultades. Las primeras iglesias misioneras eran construcciones de gran sencillez, constituidas por una amplia nave rectangular en cuyo frente oriental se adosaba un pequeño presbiterio. La nave estaba destinada a la comunidad, cuyos miembros generalmente permanecían en pie durante las ceremonias, mientras el presbiterio, junto al altar, se destinaba al clero.

En fechas recientes se han excavado algunas de las iglesias misioneras de Noruega. Los restos encontrados sugieren que muchas de las iglesias, constituidas por nave y presbiterio, se erigieron por el antiguo método de construcción por empalizadas. Este método consistía en clavar en el suelo troncos de árboles, divididos longitudinalmente en dos, unos junto a otros, para formar los muros. En el sistema más refinado, llamado de construcción sobre pilotes, los soportes de los ángulos se hundían en tierra y se colocaban entre ellos unas vigas horizontales apoyadas en el suelo. Sobre estas vigas se levantaban los muros de tablazón.

La construcción sobre pilotes suponía un avance con respecto al método de empalizada, pero los elementos de base, al estar en contacto con el suelo, podían pudrirse. Esta es la razón por la que se conservan tan pocos edificios de este tipo, a pesar de que las excavaciones arqueológicas señalan que eran frecuentes en toda Europa. Una innovación posterior aumentó la vida de las vigas de base y permitió estructuras más amplias, principalmente en los edificios con cimentación completa sobre el nivel del suelo. Sobre una hilada de losas de piedra se colocaba una armadura rectangular, con gruesos elementos llamados soleras. Los soportes verticales principales se colocaban en los ángulos de la armadura, donde las soleras se entrecruzaban.

Las tablas que constituían los muros se insertaban verticalmente en unas hendiduras de las soleras. En ocasiones, el estrecho espacio existente entre

los maderos se cerraba con unas junturas machihembradas. En otros casos se cubría con un delgado junquillo de madera. Así, los grandes soportes angulares son pies derechos, mientras la tablazón vertical de los muros recuerda las duelas de un barril. Este sistema constructivo, con la tablazón de los muros dispuesta verticalmente, permanece en pie, en contraste con el método noruego habitual de construcción en madera, llamado *laft*. En éste, la tablazón se coloca horizontalmente y se une en los ángulos mediante ensambles dentados.

Tras la construcción de las primeras iglesias de pies derechos, a mediados del siglo X, el modelo fue perfeccionándose durante varios siglos. La iglesia de Urnes, en el fiordo Luster, un brazo del fiordo Sogne, es un ilustrativo ejemplo de este proceso de desarrollo. Construida alrededor del 1150, se considera la iglesia de pies derechos más antigua de cuantas se han conservado en Noruega. Los descubrimientos realizados en 1956, durante la excavación efectuada bajo el pavimento del templo, señalan, sin embargo, que existieron antes dos iglesias en el mismo emplazamiento.

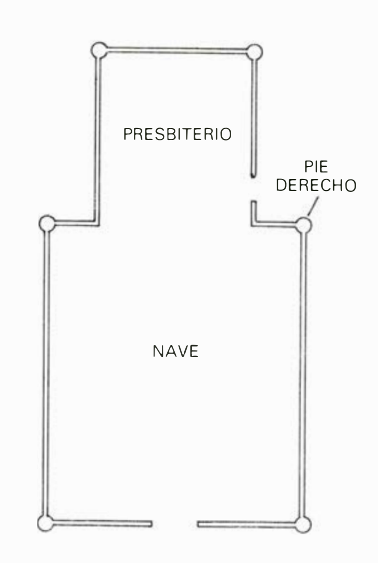
La primera iglesia de Urnes seguramente se edificó durante la transición al cristianismo. Se construyó según el método de empalizada, con todos los elementos estructurales clavados en el suelo. La segunda iglesia se levantó en la segunda mitad del siglo XI, por el procedimiento de pilotes. Tanto ésta como la anterior eran pequeñas estructuras que repetían el esquema de nave y presbiterio.

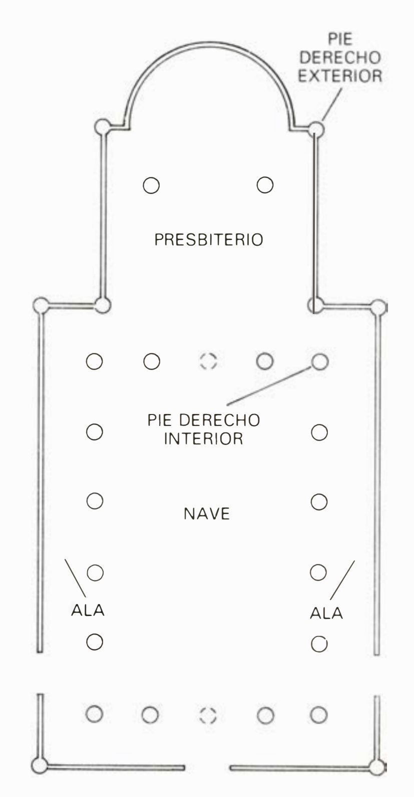
A mediados del siglo XII se derribó la segunda iglesia de Urnes, empleándose algunos de sus elementos para la construcción de la tercera iglesia, la que aún permanece en pie. Así pues, en menos de siglo y medio se construyeron en Urnes tres iglesias. El esquema de la tercera ofrece una decisiva novedad técnica con respecto a sus predecesoras. La sección central de la nave está elevada sobre el resto del edificio mediante un entramado compuesto por dieciséis grandes pies derechos que determinan la nave y las alas circundantes. El tejado del compartimento central es casi dos metros más alto que el de las alas. El llamativo diseño de la tercera iglesia de Urnes se divulgó por Noruega y se adoptó en la construcción de muchas iglesias de pies derechos posteriores.

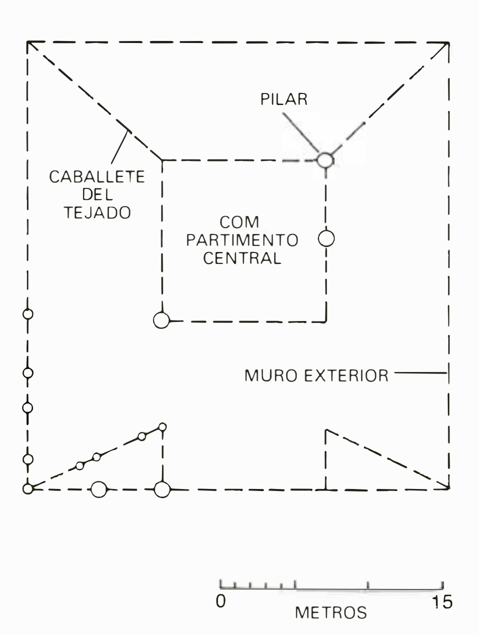
Sobre el nuevo modelo parecen haber ejercido un destacada influencia las catedrales de tipo basilical que se construían por aquellas fechas en otros lugares de Europa. La mayoría de esas catedrales se edificaron en piedra, pero es posible que los constructores noruegos vislumbraran una forma de adaptar el esquema a la madera. La mayoría de estas catedrales presentan en un extremo un profundo ábside semicircular, estructura en la que se aloja el presbiterio. La larga nave central está separada de las laterales mediante dos hileras paralelas de columnas. Los constructores de las iglesias de pies derechos tenían ejemplos basilicales a mano, pues catedrales de piedra según este modelo se construyeron en Oslo, Hamar, Stavanger, Bergen y Trondheim. Dada la similitud con las catedrales basilicales, llamaremos, al tipo de iglesia de pies derechos que aparece por vez primera en Urnes, iglesia basilical de pies derechos.

Aunque las catedrales basilicales bien pudieran haber tenido una influencia decisiva sobre los constructores de la iglesia de Urnes, existían otros modelos, culturalmente más próximos. Kenneth J. Conant, de la Universidad de Harvard, señala que las excavaciones en Gamla Uppsala, en Suecia, han sacado a la luz los restos de un templo pagano con un esquema muy semejante al de una iglesia basilical de pies derechos. Según Conant, en el templo "había un compartimento central cuadrado con pilares, en los ángulos, de un diámetro superior a los sesenta centímetros y un soporte entre [cada par de vigas grandes]. El muro exterior presentaba soportes más pequeños y [era] relativamente bajo, mientras los grandes pilares angulares determinaban una especie de torreón en el espacio cuadrado central". El templo de Gamla Uppsala se construyó a mediados del siglo XI.

Excavaciones efectuadas en Lund, Suecia, han descubierto los restos de algunas iglesias prácticamente coetáneas. Estas tempranas iglesias se levantaron siguiendo el método de empalizada.







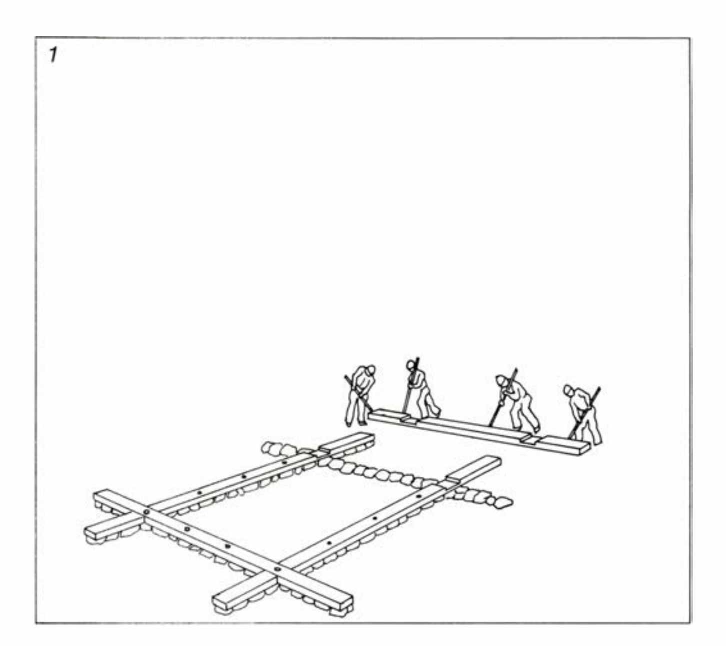
3. CONJUNTO DE PLANTAS que muestra el desarrollo de las iglesias basilicales de pies derechos, la primera de las cuales se construyó en torno al año 1100. Con anterioridad al siglo XII, se formaban con un sencillo esquema de nave y presbiterio (*arriba*). Largas soleras situadas sobre una base de losas de piedra constituían los cimientos. Sobre los ángulos se situaban seis grandes pies derechos, entre los que se colocaban los paneles que constituían los muros. En las iglesias basilicales, los soportes interiores que configuraban el elevado compartimento central se asentaban sobre cuatro grandes soleras que formaban una plataforma rectangular (*centro*). Se aprecia en este diseño cierta influencia de los esquemas de los templos paganos, como el de Gamla Uppsala, Suecia (*abajo*). El plano de este templo se ha reconstruido basándose en los pocos soportes que en la excavación se encontraron intactos (*círculos*). El templo tenía unos muros exteriores bajos y un compartimento central en forma de torreón, que presentaba en los ángulos unos soportes cuyo diámetro superaba los 60 centímetros.

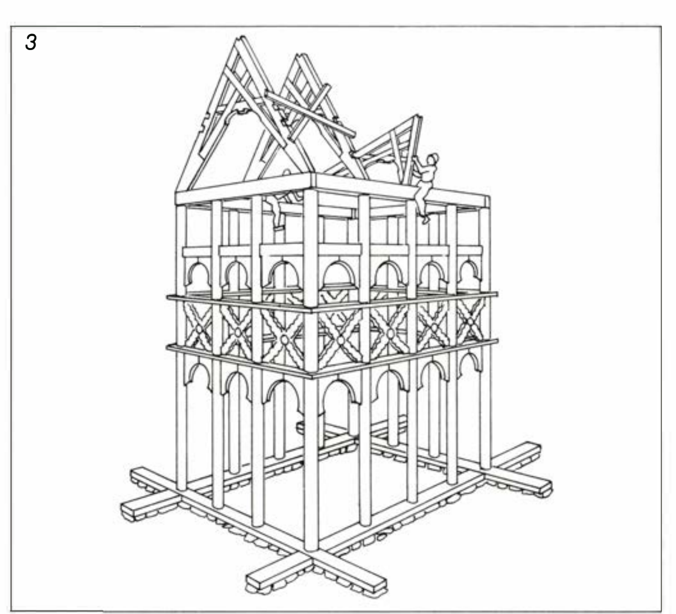
Algunas de ellas también presentan pilares interiores, como ocurre en las iglesias basilicales de pies derechos. Por ello, estas viejas iglesias misioneras suelen considerarse los precedentes de las iglesias de pies derechos.

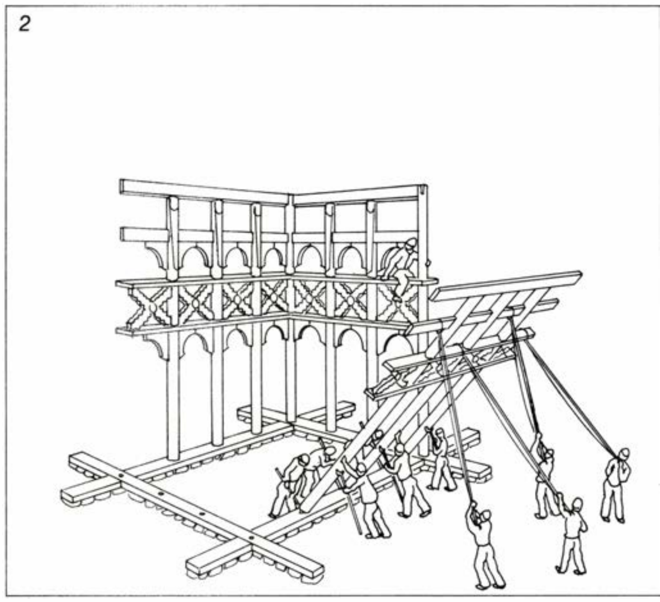
Parece, por tanto, que para la conversión de las sencillas iglesias de nave y presbiterio en los grandes modelos basilicales, los constructores de las poblaciones noruegas se sirvieron tanto de fuentes cristianas como paganas. Tal conclusión se ve reforzada por la presencia de motivos decorativos precristianos en los exteriores de las iglesias de pies derechos. Muchas de las estructuras exteriores están adornadas por minuciosos grabados de dragones y otras figuras zoomórficas, en una exuberante maraña. Los motivos ornamentales tienen variada procedencia, incluyendo las piedras grabadas con caracteres rúnicos de Escandinavia y cierto tipo de manuscritos iluminados, que en aquella época se realizaba en Irlanda. Cualquiera que sea el origen de tales grabados, poseen un claro espíritu pagano. Sólo después del año 1100 los símbolos cristianos, tales como zarcillos de vides, hacen aparición entre los dragones. Además de los grabados zoomórficos, las iglesias de pies derechos tienen en el gablete del tejado cabezas de dragones muy parecidas a las que figuraban en las proas de las naves vikingas. Estas son con frecuencia mayores y más impresionantes que las cruces que rematan el tejado.

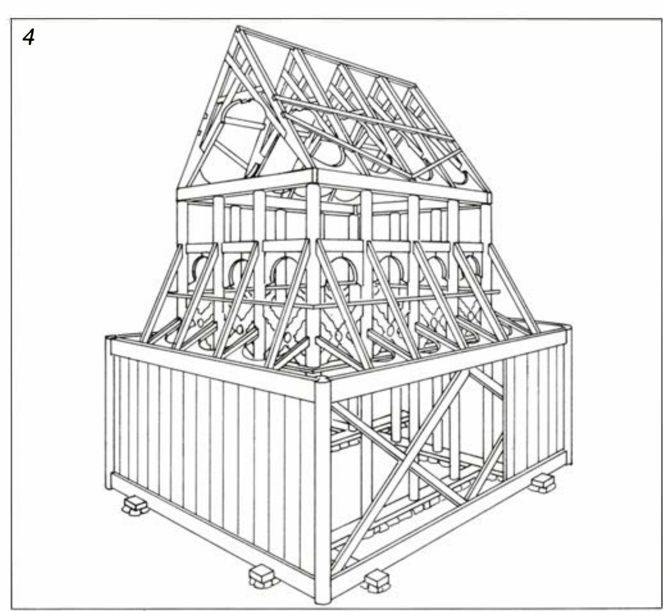
Las iglesias de pies derechos gozaron de indudable popularidad entre los habitantes de Noruega. Se estima que en el año 1300 existían en el país alrededor de 1000 de esas iglesias; muchas de ellas se construyeron en el área norte de Bergen, por los pobladores del fiordo Sogne, pero también existieron iglesias de ese tipo desde la costa sur hasta Nidaros (la moderna Trondheim). Más aún, aunque no han sobrevivido iglesias de pies derechos en Suecia y Dinamarca, recientes excavaciones sugieren que tales iglesias también se construyeron en ambos países.

Como es habitual en los edificios religiosos cristianos, la mayoría de las iglesias de pies derechos presentaban su nave en sentido este-oeste, ocupando el presbiterio el extremo oriental. Las iglesias se construían generalmente sobre lugares visualmente prominentes, y









4. CONSTRUCCION DE UNA PEQUEÑA IGLESIA BASILICAL de pies derechos, que se realizaba con métodos similares a los empleados para erigir una casa prefabricada. Las soleras se colocaban sobre un lecho de losas de piedra unidas por mortero (*1*). Las hileras de soportes del compartimento central se levantaban sobre las soleras con ayuda de puntales, y se les daba una sujeción temporal (*2*). Se afilaba la espiga de la base de los pies derechos para que encajara con facilidad en la caja. Mediante las arquerías se unían en los ángulos los grupos de soportes. Tras levantar los cuatro frentes del com-

por consiguiente quedaban expuestas por completo a los efectos del sol, el viento, la lluvia y la nieve. Para construir grandes edificios de madera en tales lugares había que superar un número considerable de problemas técnicos. La cimentación debía ofrecer una base estable al edificio y se debía asegurar la protección de los elementos inferiores contra la pudrición; los empujes verticales generados por la nieve y la propia masa de la estructura debían transmitirse a los cimientos para asegurar la estabilidad del edificio; la estructura debía soportar los importantes empujes laterales ejercidos por los fuertes vientos. Se debía prever la protección de las junturas, expuestas a la intemperie, frente al deterioro y el reemplazo de las partes que pudieran pudrirse. La forma en que estos problemas fueron solventados se deduce del análisis que hemos realizado de dieciséis de las iglesias de pies derechos que sobreviven. Las iglesias basilicales de pies derechos son las más complejas del grupo y, en lo que sigue, nos centraremos en ellas.

Los principales pies derechos interiores de una iglesia basilical están soportados por cuatro elementos horizontales llamados soleras. Estas se sitúan sobre hiladas de losas de piedra, colocadas sobre el suelo, y se unen en los ángulos por medio de ensambles a media madera. Se colocan en paralelo, por parejas, determinando al cruzarse un cuadrado con sus lados prolongados [*véase la figura 4*]. Las vigas se dentaban en sus extremos para ajustar las soleras de los muros laterales. Según fuera el diseño de la iglesia se podían añadir otras vigas a la cuadrícula determinada por las soleras.

Los cimientos de la nave y del presbiterio se construían a menudo con independencia unos de otros. La interacción de los componentes estructurales resultantes de este sistema permitía que los empujes se transmitieran, de forma complicada, a través de todo el edificio. Tal configuración tiene la clara desventaja de introducir grandes presiones en la estructura si uno de los bloques hace más asiento que otro. No obstante, nuestra investigación demuestra que no existe prácticamente tal "deformación de asiento" en las iglesias de pies derechos. Por consiguiente, parece que las condiciones del terreno, en las áreas en donde sobreviven iglesias de este tipo, son favorables para tales edificios.

Los principales elementos de la estructura que transmiten los empujes del tejado y de los muros a los cimientos son los pies derechos. Estos, cuando están situados en el interior, pueden alcanzar los 11 metros de altura, variando su diámetro desde los 30 hasta los 40 centímetros. Una espiga, el saliente que debe ajustarse en una caja, une la base del pie derecho con la solera. Sobre la parte superior del pie derecho se dispone una nueva solera, constituida por dos pequeñas vigas. El rectángulo que determinan estas soleras de la nave constituye la base del tejado de la iglesia. Por consiguiente, el compartimento central de una iglesia basilical de pies derechos tiene la forma de un paralelepípedo. El lado superior del mismo viene marcado por las soleras de la nave, mientras la base la determinan las soleras de cimentación. Los cuatro flancos del compartimento están constituidos por los pies derechos interiores, generalmente de tres a cinco en cada lado.

El compartimento central rectangular debe reforzarse para asegurar su firmeza y dotársele de inflexibilidad en varios sentidos. Se aseguraban los lados para evitar que los pies derechos se derrumbaran como las fichas de dominó. La resistencia a las fuerzas ejercidas sobre la superficie de los pies derechos se logra fundamentalmente mediante vigas horizontales, llamadas dinteles, que se colocan entre 1,25 y 2 metros por debajo del remate del pie derecho. Como veremos, los dinteles cumplen diversas funciones importantes, además de la de reforzar los pies derechos. En el ángulo recto formado por el dintel y el pie derecho hay una horquilla de madera que aumenta la rigidez. El conjunto de las horquillas constituye lo que se llama la arquería.

partimento central se colocaba el tejado (*3*). Después se añadían las alas (*4*). La tablazón que constituía los muros de las alas se ensamblaba en las soleras. La estructura completa figura a la derecha. El dibujo se basó en la iglesia de Borgund, construida en torno a 1200. Está realizado desde el presbiterio, hacia la entrada principal. El compartimento central está constituido por 14 pies derechos de 6,7 metros.

Próxima a ésta, en muchas iglesias de pies derechos hay un conjunto de tirantes realizados mediante la intersección diagonal de dos elementos situados entre cada pareja de pies derechos. Tales conjuntos, llamados cruces de San Andrés, aparecen por primera vez en torno a 1200. Numerosas iglesias erigidas con anterioridad a esta fecha fueron reforzadas con cruces mucho después de su construcción. Las cruces se destinaban probablemente a reforzar horizontalmente el conjunto de pies derechos. No obstante, en otras iglesias las cruces aparecen demasiado sueltas para que opongan alguna resistencia. De hecho, ascendiendo por el interior de la iglesia de pies derechos de Torpo, uno de nosotros (Sack) estuvo a punto de lesionarse cuando una cruz casi se desprendió de sus ensambles. En algunos pies derechos muy largos se colocaba una segunda arquería bajo las cruces. Las horquillas y las cruces se unían a los pies derechos y a las vigas horizontales mediante espigas redondas de unos 25 milímetros de diámetro. Algunas de las espigas presentan un engrosamiento terminal, aunque en otras ocasiones las espigas se introducían mediante una cuña de madera clavada en su extremo exterior.

A medida que los constructores de las iglesias de pies derechos fueron adquiriendo experiencia, empezaron a cambiar la configuración de las mismas con el fin de ensanchar la nave, facilitando el acceso y la visibilidad del presbiterio desde la nave. Esto se hizo acortando las proporciones de algunos pies derechos. Los correspondientes a los ángulos del compartimento central se dejaron intactos, pero algunos de los emplazados entre ellos se acortaron hasta casi la mitad, de tal suerte que la base de los pies derechos no descansaba sobre los cimientos, sino sobre un arco de madera que formaba una amplia horquilla. Esta característica se introdujo con gran atrevimiento. En algunas de las primeras iglesias se acortó un pie derecho del flanco este y otro del oeste. En la iglesia de Heddal, levantada en torno a 1300, alternan los pies derechos recortados con los de mayor tamaño; en la de Hurum y en la de Lomen sólo los pies derechos angulares alcanzan el suelo.

Seguramente tales modificaciones se introdujeron tras adquirir una gran experiencia en la construcción con pies derechos. El recorte de éstos habla de una gran seguridad en la fortaleza de la estructura. Además, las modificaciones hicieron necesarias una serie de horquillas que transmitiesen la carga de los pies derechos recortados a los ángulos, en donde los normalizados las transmitan a los cimientos.

Los constructores debieron levantar el compartimento interior de las iglesias pequeñas de pies derechos del mismo modo en que se erige hoy el armazón de una casa de madera. Cada conjunto de tres a cinco pies derechos, correspondientes a una de las paredes del paralelepípedo central, se situaban juntos sobre el suelo, junto al emplazamiento del edificio. Colocadas las soleras, la espiga de la base del pie derecho se insertaba en la caja de la solera. La espiga se rebajaba para que el pie se deslizara con facilidad en su caja al levantarlo. El conjunto, ensamblado en el suelo, se levantaba entonces con ayuda de puntales.

El proceso se repetía hasta levantar los cuatro lados del compartimento central. Los grupos de pies derechos se unían en los ángulos mediante las horquillas. Además, los ángulos solían reforzarse a la altura de los dinteles mediante una escuadra horizontal, que daba al compartimento central la fortaleza necesaria para resistir las deformaciones horizontales. El compartimento central de una iglesia de pies derechos es, geométricamente, similar a una caja de cartón sin base ni tapa. Si un fuerte viento incide en ángulo sobre la iglesia, el compartimento podría derribarse, de la misma forma en que, en los supermercados, las cajas de cartón suelen guardarse aplastadas antes de amonto-



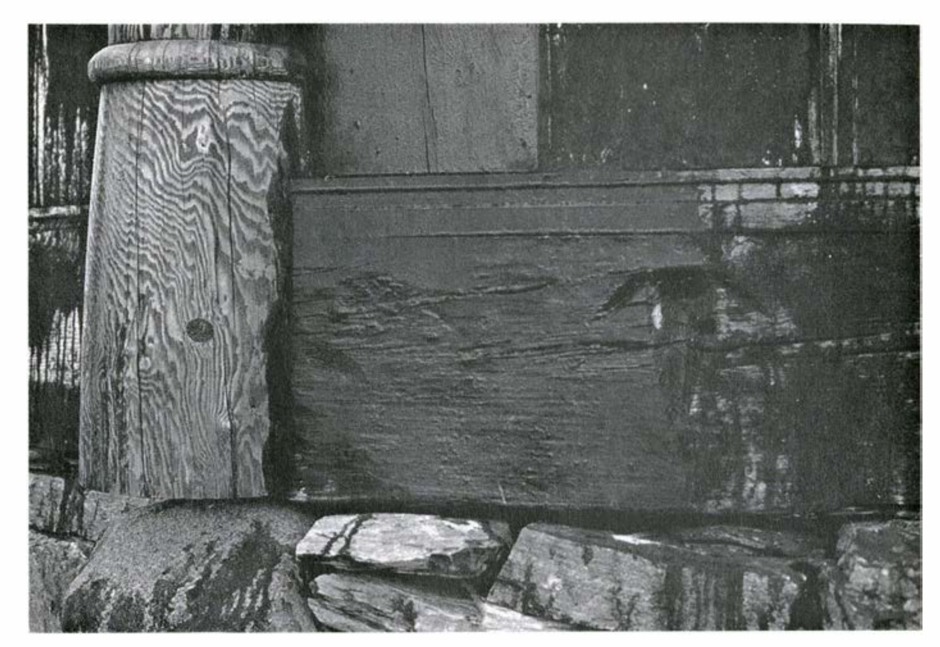
5. LA IMAGINERIA PAGANA domina la decoración labrada en el exterior de muchas de estas iglesias, tal y como se observa en este detalle de la iglesia de Hopperstad, construida en torno al 1180. Los motivos zoomórficos se enlazan con temas extraídos de las piedras escandinavas grabadas con carácteres rúnicos y de los manuscritos iluminados irlandeses. Entre otros motivos se incluyen el dragón y el grifo. Los temas de origen cristiano, tales como los zarcillos de vides, se insertan entre los elementos precristianos.

narlas. Las escuadras contribuyen a dar la resistencia necesaria.

Tras levantarse el compartimento central se colocaba el tejado, que en este tipo de iglesias es muy agudo, presentando normalmente una elevación de 3/2, o un ángulo de 56 grados. Sus pares se enlazan mediante un par de vigas, que se cruzan bajo el vértice del techo. Tal enlace en forma de tijeras tiene, a menudo, una elevación de 1/1, o un ángulo de 45 grados. Una viga horizontal, como un nudillo, se sitúa entre los pares, casi en su punto medio. Este normalmente está unido tanto a los pares como a las vigas en tijera. Pequeños elementos secundarios, llamados cerchas, que van paralelos al caballete, transmiten las cargas del tejado a los pares y a toda la armadura. Para evitar el derrumbamiento de la estructura, en las iglesias de pies derechos más modernas se incorporan unas armaduras diagonales en el plano del tejado.

El conjunto de los pares, vigas en tijera, abrazaderas y nudillos parece una auténtica estructura de armadura, pero no funciona como tal. Un tipo de ensamble muy frecuente en edificios europeos de madera es el conocido por armadura alemana o armadura en tijeras. Consta de nueve elementos: cuatro diagonales en la cima, dos más bajos unidos, dos puntales interiores y uno vertical [*véase la figura 7*]. Como en una estructura de armadura, los elementos están en tensión o en compresión. Bajo un empuje vertical hacia abajo, los cuatro elementos superiores y los dos puntales interiores se hallan en compresión y, en tensión, los dos más bajos y el vertical.

Sin embargo, en las vigas en tijera del tejado de una iglesia de pies derechos, falta, obviamente, el elemento vertical. En su lugar aparece el nudillo. Esta circunstancia hace que la estructura no funcione como una armadura ideal. Un análisis de la misma muestra que muchas vigas están sometidas al combado, a la tensión y la compresión, a la vez. No es soprendente que los constructores noruegos del siglo XII no construyeran una armadura ideal. El diseño de ésta no se formalizó hasta principios del siglo XVI; con anterioridad a esa fecha los problemas técnicos se resolvían por el método de tanteo. No obstante, el nudillo cumple una función importante: redistribuir las fuerzas en el conjunto y con ello minimizar los componentes horizontales de las fuerzas distribuidas a las paredes. Si no existiese el nudillo, la parte superior de los muros podría desplazarse bajo el peso de la nieve.



6. DRENAJES DE LAS SOLERAS; ayudan a proteger de la pudrición los vulnerables cimientos de las iglesias. A la izquierda de esta imagen de los cimientos de la iglesia de Øye se encuentra la base de uno de los pies derechos exteriores de los ángulos. El largo elemento horizontal es la solera; sobre ella se inserta la tablazón que constituye los muros de las alas. El agua que corre por ellos puede concentrarse en la unión entre la tablazón y la solera. Para sacar el agua de los ensambles, los constructores abrieron en la solera pequeños agujeros, separados entre sí unos cuantos metros. La salida de uno de tales drenajes se observa a la derecha. Gracias a esta práctica, los cimientos de la mayoría de estas iglesias se encuentran en excelentes condiciones. Como protección adicional, las iglesias solían embrearse cada dos o cinco años.

Colocados ya los pares y los tirantes del tejado, se añadían las planchas de madera de la cubierta, empezando por el caballete y acabando por los aleros. Estas planchas se colocaban paralelas al caballete. Las dos series de planchas se unían, generalmente, mediante clavos de madera. En algunas iglesias se cubrían por una tercera serie, realizada por duelas de madera o tablas de ripia sin labrar. Las duelas se cortaban de forma peculiar, lo que daba al tejado la apariencia de las escamas de los peces.

Tras instalar el tejado se añadían las alas. Los tejados de las alas, por ser más bajos que los del compartimento central, resaltaban las líneas verticales del edificio y daban al exterior un perfil parecido al de una serie de escalones inclinados. La parte superior del faldón que soportaba el tejado a un agua se unía al pie derecho mediante espigas de madera. El extremo inferior descansaba en un rebajo de la viga horizontal, llamada solera, en que remata el muro lateral.

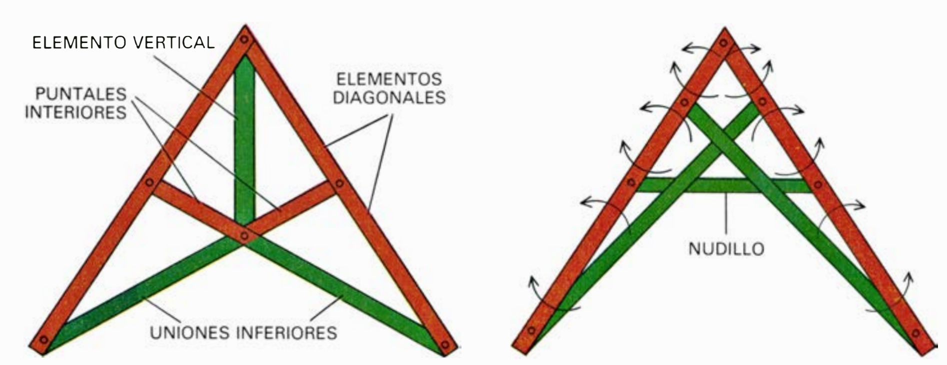
Bajo los pares del ala hay una cruceta con una inclinación menor que la de los pares. Llamada tornapunta, está unida con espigas tanto al muro lateral como al pie derecho interior, por lo que puede transmitir a la vez fuerzas de tensión y de compresión. Por el contrario, los pares de las alas están unidos sólo al pie derecho interior. El otro extremo de los pares de las alas, donde se encuentra con la solera, sólo descansa en un rebaje de la misma. Así pues, sólo transmite fuerzas de compresión.

Entre los tornapuntas de las alas hay una gran abrazadera curva conocida como cuadrante. Enlazadas con la solera, ajustan notablemente la estructura del alero, circunstancia de la mayor eficacia a la hora de resistir las fuerzas horizontales, paralelas al eje de la nave.

Uno de los procesos finales de la construcción era la colocación de la tablazón vertical de los muros, entre la solera inferior y la situada sobre los muros laterales. Ambas se acanalaban para ajustar las espigas de los extremos de los tableros. La ranura existente entre las tablas se cerraba con una juntura machihembrada o con un junquillo.

La existencia de unas alas flanqueando la nave central contribuía de manera especial al fortalecimiento lateral de toda la iglesia. Consideremos que un viento sopla en ángulo recto sobre el eje de la nave; los empujes del mismo sobre los muros laterales se transmitirán de la tablazón a la solera y al pie derecho interior por medio de los tornapuntas y los cuadrantes. La estructura tenderá a deformarse en la dirección del viento. En la deformación, el par de barlovento tiende a responder como un elemento de tensión. Puesto que el par está unido solamente a uno de sus extremos, no sirve para transmitir las cargas.

Los empujes laterales se transmiten a



7. ELEMENTOS DEL TEJADO de una iglesia de pies derechos, que constituyen una armadura modificada que funciona de manera diferente a como lo haría una armadura ideal. En ésta, los distintos elementos están en tensión (*verde*) o en compresión (*rojo*), pero no en ambas a la vez. No concurre otro tipo de fuerzas en el conjunto. El dibujo de la izquierda es la estructura conocida como armadura alemana, que funciona como una armadura ideal. En las vigas en tijera del tejado de una iglesia de pies derechos, mostrado a la derecha, el elemento vertical se sustituye por un nudillo. Por cuya razón algunos elementos están sujetos a fuerzas de alabeo (*flechas*). El nudillo ayuda a la estructura a resistir los empujes laterales.

través del compartimento central de la iglesia por medio de los pies derechos interiores, la estructura del tejado, las arquerías y las cruces de San Andrés. En la zona del edificio que queda a sotavento, los pares y las tornapuntas actúan como pequeñas armaduras para transmitir los empujes a los muros laterales y a los pies derechos exteriores. De este modo, las alas realizan una función estabilizadora mucho mayor que la de un arbotante en una catedral gótica. Sin embargo, en una catedral gótica es el volumen del arbotante lo que aporta la estabilidad. En una iglesia de pies derechos dicha capacidad está en toda la sección transversal, capaz de transmitir con eficacia los empujes de un lado del edificio a la base del lado contrario, lo cual es un factor de estabilidad.

Los constructores de las iglesias de pies derechos se valieron de una interesante combinación de elementos de viguería y estructura de armadura para lograr la resistencia contra los empujes verticales y horizontales. El tejado es una estructura de armadura modificada. El compartimento central, que debe resistir tanto empujes verticales como horizontales, está realizado fundamentalmente por elementos de viguería. Las alas, que deben resistir importantes empujes laterales, actúan cual pequeñas armaduras. La combinación de ambos sistemas es muy eficaz.

Los materiales empleados en una iglesia de pies derechos son casi tan importantes como la estructura misma. Los constructores parecen haber sido muy hábiles artesanos en la selección y preparación de la madera. Los pies derechos, la tablazón y muchos otros elementos estructurales son de pino albar (*Pinus silvestris*), abundante en Noruega. Se tuvo un especial cuidado en la selección y secado de los pies derechos principales. De los muchos árboles a su alcance, los constructores seleccionaban unos cuantos y les cortaban las copas, sin arrancarlos. Se dejaba secar el tronco enraizado durante cinco u ocho años, tras lo cual se talaba y se cortaban a medida los elementos estructurales de mayor envergadura. Las zonas exteriores del tronco, más ricas en savia, se eliminaban, empleándose solamente el corazón, más duradero, para los pies derechos. La madera de muchas de las piezas curvas eran de una especie noruega de abedul (*Betula verrucosa*). La madera para la mayoría de las espigas y otros elementos de ensamble eran de enebro común (*Juniperus communis*), una madera compacta y flexible.

Nuestras investigaciones sobre las iglesias de pies derechos demuestran que las prácticas de secado de los constructores medievales les proporcionaban mejor material del que disponen los constructores de hoy en día. En la mayoría de las iglesias que visitamos, los grandes pies derechos interiores se encontraban en buenas condiciones. En la iglesia de Urnes, un pie derecho de 800 años aún rezumaba resina. La madera de los pies derechos originales estaban en buen estado y casi sin grietas. En realidad, las peores grietas se daban en dos pies derechos que habían sido colocados hace unos 40 años. Ambos presentaban grietas de casi 40 milímetros de ancho que recorrían la superficie desde la base hasta la cima. Como era de esperar, los pies derechos exteriores solían estar más deteriorados que los interiores, pero incluso los que se encontraban a la intemperie aparecían prácticamente sanos.

El especial diseño de estas iglesias les favorece en su resistencia al fuego. Los elementos de mayor tamaño, tales como los principales pies derechos de la estructura, pueden sobrevivir al fuego, porque son de combustión lenta. Así, una iglesia de pies derechos podía resistir el paso de un incendio sin llegar a hundirse y, para repararla, no había más que sustituir los elementos carbonizados. Sin embargo, si el fuego era lo bastante intenso, también los elementos estructurales de mayor envergadura se quemaban. Se sabe que un buen número de iglesias de pies derechos fueron destruidas por el fuego. Está en vías de ejecución un programa de protección de todas las iglesias de este tipo mediante la colocación de sistemas de aspersión automática.

La resistencia de las iglesias de pies derechos a la acción del clima y del tiempo se ha visto favorecida por la existencia de numerosos detalles estructurales. A lo largo del dintel y de la cara exterior de la solera, el peligro de putrefacción es grande. El agua corre por los paneles de tablazón hasta detenerse en las vigas horizontales, donde al recogerse podía penetrar por las junturas. Para neutralizar el problema, los constructores practicaron pequeños drenajes a trechos de unos cuantos metros a lo largo del dintel y de la solera, para permitir la salida de agua de las ranuras en que se aloja la tablazón de los muros. La práctica ha sido eficaz; sólo hemos encontrado una solera en avanzado estado de descomposición.

En muchas iglesias una galería cubierta rodea el edificio. Tal pórtico no tiene ninguna función estructural importante, pero protege la parte baja de la iglesia de las inclemencias del tiempo. En la iglesia se coloca, alrededor de la base de los pies derechos, unas cubiertas bulbosas de madera. Las cubiertas son reemplazables; si una se pudre se puede instalar otra nueva. La mayoría de las iglesias se embrean cada dos o cinco años, práctica que les confiere un olor característico.

Dos factores de carácter más general han contribuido también a la conservación de las iglesias de pies derechos. En primer lugar, muchos de los elementos estructurales están sometidos a bajos esfuerzos. En la mayoría de los edificios que se levantan en la actualidad, los elementos de soporte están sometidos a presiones que equivalen a una fracción considerable de su resistencia máxima. Con ello se ahorra dinero y se reduce el peso y la sección de los mismos. Por el contrario, nuestros cálculos demuestran que, en una iglesia de pies derechos, éstos soportan presiones que vienen a ser sólo la décima parte de lo que pueden aguantar. Tan gran reserva de posibilidades significa que es improbable que los elementos puedan fallar por exceso de carga. En segundo lugar,

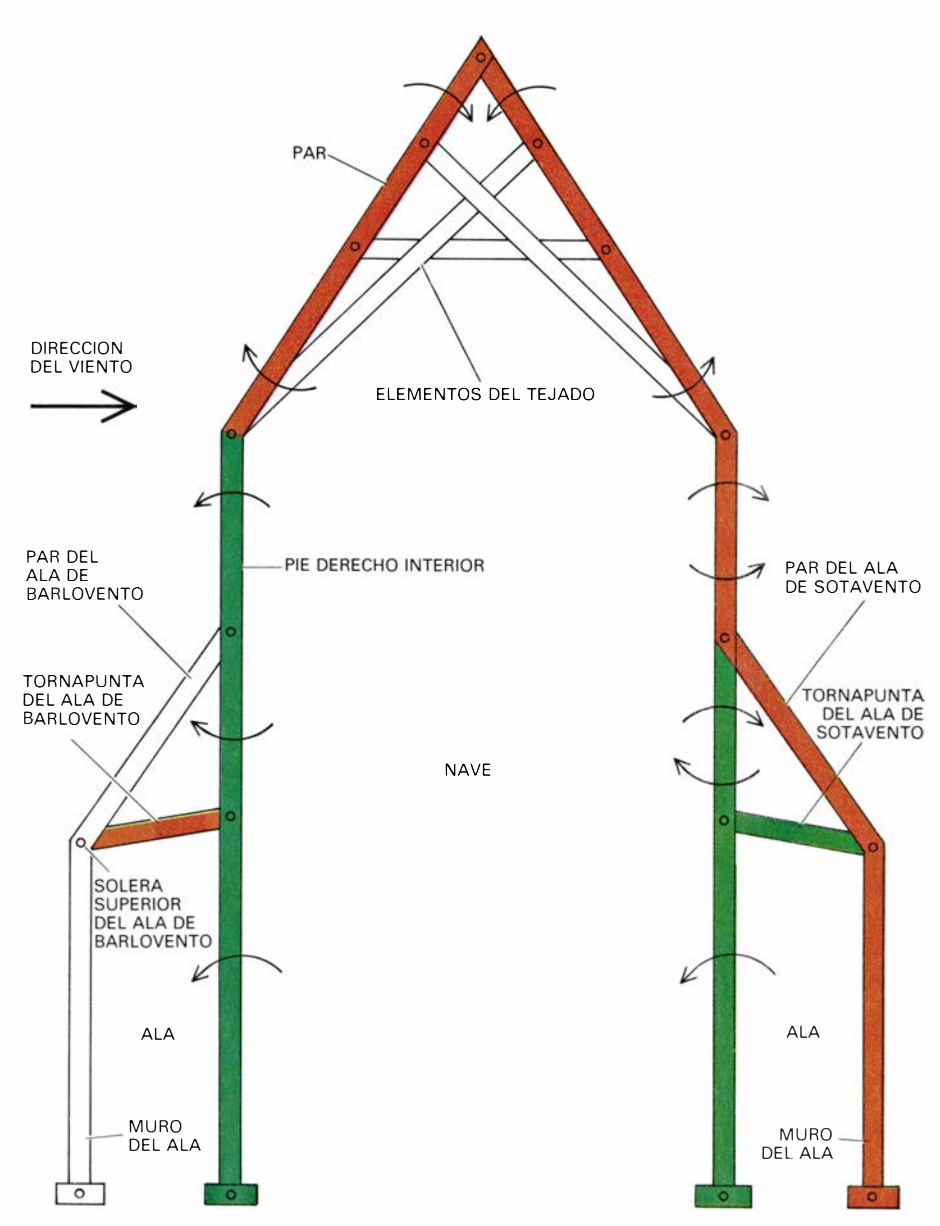
el modelado de todas las espigas y de los elementos de ensamble del edificio es muy preciso. Como hemos señalado, la capacidad de la sección transversal de una iglesia de pies derechos para transmitir empujes lateramente es un factor importante para su integridad estructural. Sólo si el vaciado, entallado y modelado de los elementos de ensamblaje es preciso, los empujes pueden transmitirse con eficacia sin dañar el edificio.

Con el paso del tiempo, las iglesias de pies derechos han resistido con más éxito los derrumbamientos y los fallos estructurales que las amenazas derivadas de los cambios de necesidades y de gustos de las comunidades donde se construyeron. Alrededor de 1350, cuando el número de iglesias era más elevado, la peste bubónica asoló Noruega, al igual que otros países europeos. La población rural se redujo enormemente y la construcción de iglesias se paralizó. Al recobrarse, la influencia correspondía a nuevas actitudes religiosas; la reforma se introdujo en Noruega en 1536. Las iglesias de pies derechos se consideraron entonces demasiado pequeñas, oscuras y frías. Muchas de ellas se sustituyeron o remodelaron. Otras no se conservaron adecuadamente y se arruinaron. En 1800 sólo quedaban unas 100.

En 1814, Noruega, que había formado parte de Dinamarca durante 400 años, se hizo independiente. La independencia se vio acompañada de un renacer del orgullo nacional y de un interés por la religión. En 1851, una ley señaló la necesidad de que todas las iglesias pudieran acoger al 60 por ciento de la comunidad, lo que llevó a la destrucción de muchos viejos edificios. Al mismo tiempo, como contrapartida, se suscitó un gran interés por la herencia cultural del país y algunos noruegos se empeñaron en la conservación de las iglesias de pies derechos que habían sobrevivido.

Uno de los dirigentes de esta operación de salvamento fue el pintor Johan Christian Dahl. Cuando la iglesia de Vang, en el pueblo de Valdres, estaba a punto de ser derribada, Dahl realizó un importante esfuerzo para conseguir los fondos necesarios para comprarla. Aunque fracasó en su intento, persuadió al rey Federico Guillermo III de Prusia para que comprase la iglesia, que se desmanteló y reconstruyó en un pequeño pueblo de Silesia (actualmente territorio polaco). Allí sigue en pie la iglesia. El episodio despertó un enorme interés entre los noruegos y en 1844 se fundó la Sociedad para la Protección de Monumentos Antiguos de Noruega, destinada a la conservación de las iglesias de pies derechos existentes. Subsisten en la actualidad 29 de ellas, muchas profundamente alteradas o trasladadas de su emplazamiento original.

Entre las iglesias de pies derechos que permanecen en pie sin modificar y en su primitivo emplazamiento, están, sin duda, dos de las más hermosas, la iglesia de Urnes, la que mejor representa la génesis del tipo basilical, y la iglesia de Borgund. Su supervivencia, junto con los resultados de nuestras investigaciones, demuestran que las iglesias de pies derechos estaban bien diseñadas y construidas. Confirman el hecho de que se pueden construir estructuras de madera de vida indefinida, siempre que se cumplan una serie de condiciones. La madera debe seleccionarse y secarse cuidadosamente; se debe poseer una meticulosa práctica en el diseño; se debe prestar la atención necesaria a los detalles que pudieran anticipar la ruina; los métodos de construcción deben ser de alta calidad y la estructura debe conservarse continuamente, para reducir al mínimo el deterioro. Si se satisfacen estas condiciones, los edificios de madera resisten el tiempo.



8. RESISTENCIA A LOS EMPUJES LATERALES en una iglesia de pies derechos; se logra gracias a que toda la sección transversal de la misma funciona como un conjunto único. En este dibujo correspondiente a una sección transversal de la nave, los elementos estructurales en tensión aparecen en verde y, los que se hallan en compresión, figuran en rojo. Las fuerzas de alabeo se indican por flechas. Un fuerte viento sopla de izquierda a derecha. En este caso, la solera del muro de barlovento transmite los empujes a los pies derechos interiores mediante los cuadrantes y los tornapuntas de las alas. (En el dibujo se ha suprimido la arquería.) Los pies derechos, la estructura del tejado y el caballete del mismo transmiten los empujes a través del compartimento central. A sotavento, los pares de las alas y los tornapuntas funcionan como una pequeña armadura, que traslada los empujes al muro de sotavento. Así, las alas ejercen una función estabilizadora, similar a la de los arbotantes en una catedral gótica, de cuya masa depende su estabilidad. En una iglesia de pies derechos se logra mediante la respuesta de toda la sección transversal.

Tipografía digital

La mayoría de las letras de imprenta se producen por ordenador, en vez de usar metal fundido o métodos fotográficos. Las compositoras digitales crean nuevas formas de letra, a una velocidad de 15.000 caracteres por segundo

Charles Bigelow y Donald Day

Los caracteres tipográficos de una página de *Investigación y Ciencia* se han compuesto en una máquina cuya operación es digital. La página se forma sobre la pantalla de un tubo de rayos catódicos mediante un haz de electrones que barre 800 líneas verticales por pulgada (una pulgada equivale a 25,4 milímetros); el movimiento del haz es lo suficientemente rápido como para producir una página de texto en unos 15 segundos. La resolución de la pantalla es mucho mayor que la de un receptor corriente de televisión, pero en otros aspectos ambos dispositivos operan de forma análoga. Los minúsculos elementos gráficos individuales o "pixels", situados a lo largo de cada una de las líneas de barrido, brillan o permanecen oscuros, encendiendo o apagando el citado haz, a medida que transcurre la operación de barrido. La figura resultante está constituida por una serie de trazos rectilíneos verticales observables individualmente con la ayuda de una lente de aumento; ahora bien, el ojo humano percibe globalmente al conjunto bajo la forma de una página llena de letras con formas redondeadas.

Aunque para controlar la secuencia de encendidos y apagados del haz de electrones se necesita un ordenador digital, la propia letra es en sí misma digital en el sentido de que está constituida por un conjunto de elementos discretos. Estos elementos pueden ser rayas, elementos gráficos, colores, tonalidades distintas de gris o cualquier otra unidad gráfica a partir de la cual se pueda formar una letra. De aquí que la tipografía digital no sea estrictamente una novedad: las piezas de un mosaico, los elementos que forman un bordado y las filas de luces utilizadas en las marquesinas de los teatros han servido desde hace mucho tiempo para representar, más o menos toscamente, los caracteres alfabéticos. Sin embargo, estos letreros digitales se han considerado siempre meras curiosidades tipográficas y están basados en técnicas muy diferentes de las que se suelen utilizar en el diseño y composición de las letras. Una letra tradicional no es digital sino analógica; su forma final cambia suavemente si variamos con continuidad el proceso utilizado en su creación, tal como la presión de un pincel sobre un papel o el contorno de un punzón que golpea sobre la matriz o el molde que se utiliza para fabricar los tipos de metal fundido. Con el desarrollo de los ordenadores y la electrónica digital, la tipografía ha tratado de sustituir totalmente, a lo largo de los últimos 15 años, los textos analógicos por textos digitales, en un proceso comparable al que tuvo lugar durante el Renacimiento con el paso de la escritura manual a la imprenta.

Se ha estimado que en Estados Unidos se reproducen diariamente unas 10^{14} letras; evidentemente las letras constituyen una buen parte de la estructura visual de nuestra civilización. Aun cuando algunas de estas letras están destinadas fundamentalmente a ser "leídas" o decodificadas por una máquina, en última instancia todas ellas deben ser leídas por las personas si los canales de comunicación entre los escritores y los lectores permanecen inalterados. Pero el aprendizaje de la lectura es difícil y costoso. Con el fin de proteger tal inversión educativa, las letras que leen los adultos no deben ser muy diferentes de las letras que aprendieron de niños, y las letras que lee la actual generación no deben ser significativamente diferentes de las creadas por las generaciones precedentes. La transición a la tipografía digital plantea una cuestión muy sutil con implicaciones de largo alcance. ¿Cómo es posible aprovechar plenamente la tecnología digital y asegurar al mismo tiempo que las letras digitales conserven la calidad de las letras tradicionales, cuya belleza y legibilidad han contribuido de forma tan sustancial a nuestra cultura?

Las ventajas de la tipografía digital son innegables: una vez que las letras se han representado como elementos discretos, pueden codificarse como propiedades físicas discretas y diferenciadas en cualquier medio conveniente, son susceptibles de ser procesadas como unidades de información (bits) mediante un ordenador, pudiendo ser transmitidas a grandes distancias en forma de impulsos de corriente y decodificadas para recomponer las letras originales por parte de las personas que reciben el correspondiente mensaje. Una vez que un tipo se ha digitalizado, está de hecho codificado en el lenguaje binario del ordenador, y el tamaño, la forma y las restantes características de las letras pueden modificarse mediante un programa de dicho ordenador. Dado que algunas variedades de tipo digital se pueden leer con la ayuda de máquinas, el contenido semántico de la información representada por las letras

1. MUESTRAS DE LETRAS DIGITALES basadas en los diseños de la primera fila debidos a Kris Holmes y a uno de los autores de este artículo (Charles Bigelow). La resolución va disminuyendo a partir de la segunda fila hasta llegar a la última. Las letras de cada fila se han formado superponiendo una rejilla de cuadrados de tamaño cada vez mayor sobre las letras del modelo de la primera fila. Si alguna parte de la letra del modelo coincide con el centro de uno de los cuadrados de la rejilla, se ennegrece todo el cuadrado; en caso contrario, el cuadrado se deja en blanco. Dado que las letras digitales están constituidas por elementos discretos, resultan muy apropiadas para su almacenamiento, transferencia y manipulación por parte de sistemas electrónicos digitales y ordenadores. Sin embargo, esta técnica impone también determinadas restricciones al diseño tipográfico; en este sentido, las formas básicas de las letras deben estar "bien templadas", es decir, deben mantener la máxima legibilidad posible para una amplia variedad de representaciones digitales y, por consiguiente, el diseñador habrá de tener en cuenta la existencia simultánea de múltiples versiones de las imágenes de las letras. Las letras digitales que se reproducen en esta figura fueron obtenidas por la firma Autologic, Inc., de Newbury Park, California.

RNQbaeg

RNQbaeg

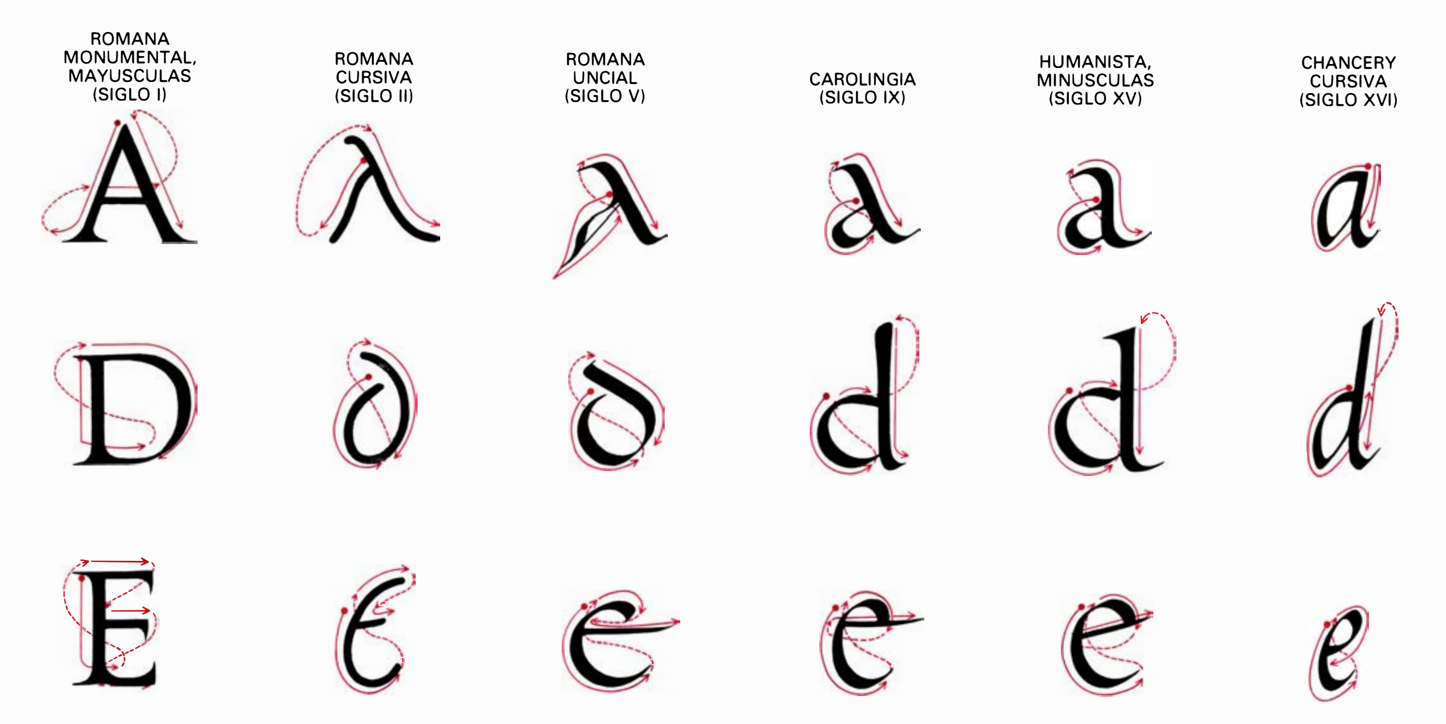
RNQbaeg

RNQbaeg

RNQbaeg

RNQbaeg

RNQbaeg

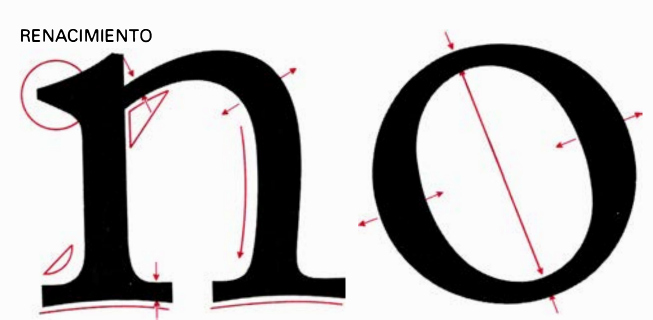


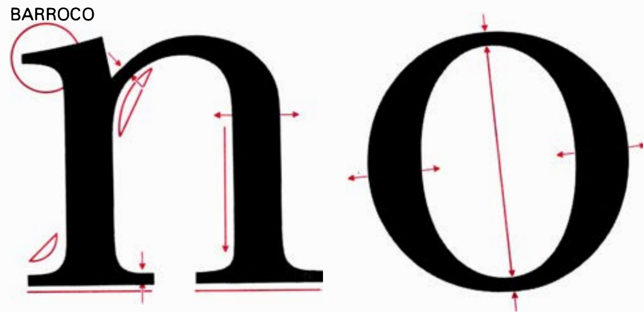
2. POR LETRAS DUCTALES se conocen las escritas a mano, cuya topología básica es el resultado de una serie de movimientos suaves de una herramienta de escritura sobre una superficie. El recorrido de dicha herramienta se denomina "ductus". El ductus de cada letra se representa en color; las líneas discontinuas indican los tramos en que el ductus es invisible en la forma final de la letra, es decir, señalan la trayectoria de dicha herramienta cuando no está

se puede manipular por un ordenador de forma parecida a como maneja las letras. Al contrario de lo que sucede con la información de tipo analógico, la información digital es muy resistente al ruido o a la degradación que se produce siempre que se transmite una señal. Los dispositivos de recepción digital sólo necesitan distinguir entre dos estados de una señal (digamos entre sí o no) con el fin de decodificar y recuperar la información transmitida originalmente.

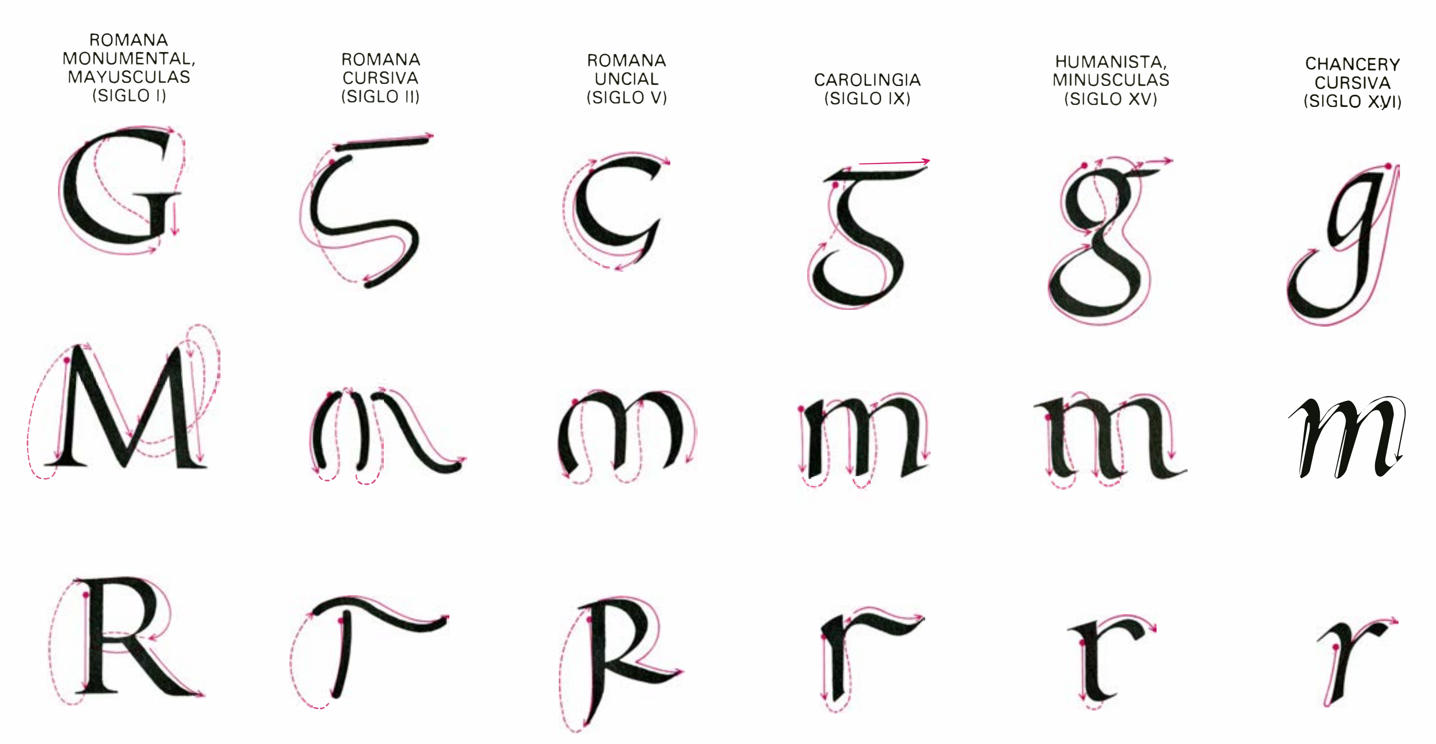
La tipografía digital se adapta a una gran variedad de dispositivos de salida. La composición de muchos periódicos y revistas de Occidente se realiza utilizando compositoras de rayos catódicos. Recientemente se ha introducido una nueva generación de impresoras láser de alta resolución en las que un haz de láser fuertemente colimado reemplaza al haz de electrones como instrumento de escritura. Estos dispositivos proporcionan una imagen tipográfica sobre una plancha de impresión obtenida directamente a partir de la información digital, o una exposición fotográfica sobre un papel o película como etapa intermedia. Las impresoras láser de baja resolución, llamadas también compositoras de texto, se utilizan como dispositivos de salida en procesos de datos y en la publicación de formularios y documentos en pequeñas cantidades. En estos dispositivos el haz del láser produce la imagen de un texto a partir de la carga electrostática depositada sobre una cinta o tambor. Partículas finas pigmentadas (partículas "toner") son atraídas a la cinta o al tambor en los lugares donde se ha creado una carga eléctrica, transfiriéndose a un papel de forma análoga a como lo hace la tinta de imprenta, y, finalmente, se funden con dicho papel por acción del calor.

Están también en servicio otros muchos dispositivos cuya resolución es bastante baja; en estos dispositivos, el número de elementos gráficos (pixels) de que se dispone para formar las letras es mucho menor que el que sería necesario para reproducir todos los detalles de las letras. Por ejemplo, la impresora de que disponen frecuentemente los ordenadores, los grandes y los personales, está constituida por una columna de finas agujas. Los extremos de dichas agujas golpean sobre una cinta impregnada con tinta y forman una figura de puntos a lo largo de una columna vertical del papel. A medida que dicha co-





3. EVOLUCION DE LAS LETRAS GLIPTALES, fiel reflejo de los cambios fundamentales que ha experimentado la tecnología de la producción de letras a raíz de la invención de la imprenta con tipos móviles. La imagen de una letra ya no es el resultado de una serie de trazos realizados "al vuelo" por un amanuense; cada letra se grababa trabajosamente en el extremo de un punzón de acero como copia patrón. Las primitivas letras gliptales, así las utilizadas durante el Renacimiento, imitaban las formas de la escritura manual y muchos de sus detalles (*color*) simulaban los efectos que produce el uso de una



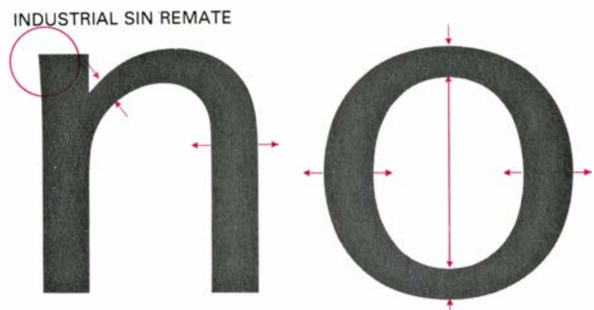
en contacto con la superficie utilizada en la escritura. Los alfabetos tales como las letras romanas mayúsculas, en los que una alta proporción del ductus es invisible, se denominan formales, en tanto que los alfabetos tales como las letras cursivas Chancery, en los que una alta proporción del ductus es visible, se denominan cursivos. La evolución del alfabeto latino es el resultado de una síntesis entre dos tendencias opuestas, la escritura formal y la cursiva.

lumna de agujas se mueve a lo largo de la página, un conjunto de electroimanes actúa sobre los extremos de las agujas y así la sucesión de figuras diferentes de puntos verticales que va apareciendo sobre la página constituye un texto digital. Se han desarrollado también unas impresoras por chorro de tinta en las que las imágenes de las letras se obtienen controlando la trayectoria de las gotitas de tinta en su vuelo hacia el papel. La trayectoria de las gotitas se controla cargándolas eléctricamente y haciéndolas pasar a través de un campo electrostático. La importancia de las denominadas copias suaves, es decir, las imágenes transitorias de tipo digital que se forman en los tubos de rayos catódicos, en las pantallas de cristal líquido y en los diodos emisores de luz ha ido aumentando con el número de ordenadores personales y los sistemas de tratamiento de textos.

La capacidad de todos estos dispositivos para reproducir caracteres tipográficos suele expresarse por el número de pixels que existen en cada uno de los lados de un cuadrado, llamado cuadrado eme, igual al tamaño de la letra de un determinado tipo. (Existen unos 72 de estos cuadrados por pulgada.) La longitud del lado de uno de estos cuadrados standard es ligeramente mayor que la distancia que separa el extremo superior de una letra tal como la *h* minúscula y el extremo inferior de una letra tal como la *y* minúscula. Para letras del tipo "pica", tipo de 12 puntos, la mayoría de las máquinas de componer digitales de alta resolución tienen una resolución del orden de 100 a 300 líneas por eme; las máquinas de componer electrostáticas, una resolución del orden de 33 a 80 líneas por eme; y los dispositivos de salida digital de baja resolución, una resolución que puede variar entre 10 y 30 líneas por eme.

La velocidad, versatilidad y bajo coste de la tipografía digital ha extendido su uso de forma irresistible. Por otra parte, si las letras se diseñan sobre un bastidor o rejilla de elementos gráficos digitales lo suficientemente fina, las diferencias entre la letra original y la letra digital son prácticamente indistinguibles. En principio, si se utiliza un bastidor digital adecuado, es perfectamente posible imitar cualquier tipo de letra tradicional impresa o manual, no importa cuán refinada sea. Sin embargo, en la práctica, la tipografía digital, como toda innovación tecnológica,





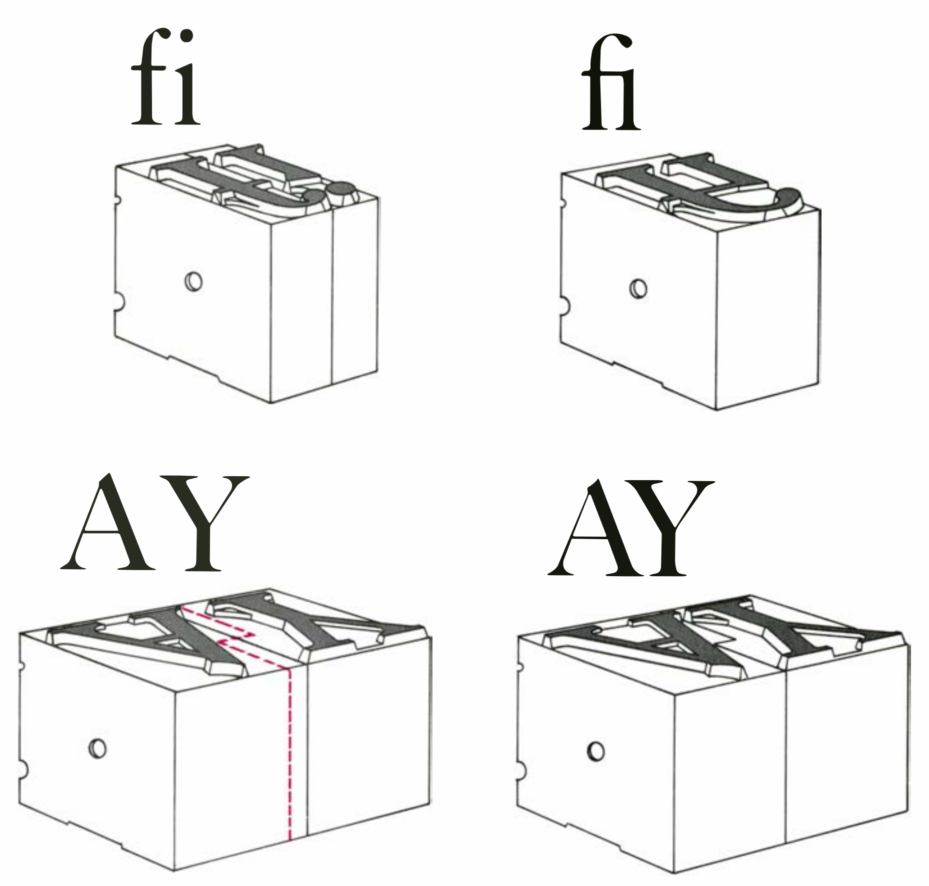
pluma gruesa cuando forma un ángulo de unos 30 grados con la horizontal. En general, el contraste entre los elementos gruesos y finos es relativamente bajo. En las versiones barroca y neoclásica, la simetría, la armonía de la estructura y el contraste entre los trazos gruesos y finos importa más que el propio trazo de la pluma. Los diseños más recientes han suprimido los remates o los han convertido en elementos independientes, como sucede con ciertos tipos industriales sin remate no representados en esta figura. Los dibujos reproducidos aquí se deben, al igual que la caligrafía de la figura 2, a Holmes.

conlleva problemas específicos y apremios selectivos.

Por ejemplo, no toda la tipografía digital puede ser de alta resolución. En primer lugar, existen determinadas limitaciones técnicas sobre el tamaño del punto que se puede reproducir utilizando métodos del estilo del chorro de tinta y las resoluciones máximas alcanzables son demasiado toscas para escapar a una simple inspección visual. Por otro lado, la velocidad y las ventajas de coste de la tipografía digital se reducen a medida que aumenta el número de bits digitales que se necesitan para formar una letra. El número de pixels por letra aumenta con el cuadrado de la resolución lineal del dispositivo de impresión: para duplicar la resolución lineal de un dispositivo es necesario aumentar en un factor cuatro la cantidad de información o número de bits a transmitir y procesar. Aunque existen métodos de cálculo para comprimir los datos existentes en el mapa de bits que configura una determinada letra, la relación general entre coste y resolución mantiene plenamente su validez.

Quizás el problema más difícil con que tiene que enfrentarse el diseñador de tipos sea lograr un uso eficaz de la enorme flexibilidad de la tecnología digital. Por ejemplo, el texto de un documento puede escribirse en primera instancia en la terminal de un tubo de rayos catódicos con una resolución de 10 líneas por eme para una letra del tipo pica. A continuación puede sacarse una prueba de ese mismo texto utilizando una impresora de matrices de agujas con una resolución del orden de 20 líneas por eme, haciéndola circular para proceder a las correcciones y comentarios pertinentes utilizando una impresora de láser con una resolución de 50 líneas por eme y, finalmente, se puede obtener una copia lista para su publicación definitiva con una resolución de 200 o más líneas por eme en una compositora de rayos catódicos. De forma análoga, cuando se pretenden presentar letras digitales en tamaños diferentes en una máquina con una rejilla fija, es necesario contar con un mapa de bits diferente para cada tamaño de letra. Evidentemente, en tales circunstancias, conviene que el diseñador de tipos cree un modelo único o pauta de letra a partir del cual se generen posteriormente todas las formas en que aparece dicha letra sobre máquinas diferentes y con distintos tamaños.

Con el fin de apreciar la magnitud de este problema, tomemos en consideración las múltiples formas de letras que



4. TIPOS METALICOS. Se funden en unas piezas cuya anchura es constante; por tanto, cuesta reproducir las variantes en las formas y en los espaciados para las diferentes combinaciones de letras, típicas en un buen manuscrito. Durante las décadas subsiguientes a la invención de los tipos móviles, algunos impresores fundieron centenares de variantes de las formas de las letras; esta práctica eliminaba casi por completo las ventajas en tiempo y economía que suponía la nueva tecnología. Más tarde, los diseños de las letras se fundieron en las piezas metálicas de manera tal que su ensamblaje para formar las palabras proporcionara automáticamente el espaciado adecuado. Ciertas combinaciones tales como la *fi* se resistían a estos diseños especiales y se fundieron para constituir un cuerpo metálico único llamado politipo. Otras combinaciones de letras, la *AY*, no se podían espaciar, a menos que los tipos se ajustaran a mano.

6 PUNTOS

RQEN baegnov

8 PUNTOS

RQEN baegnov

14 PUNTOS

RQEN baegnov

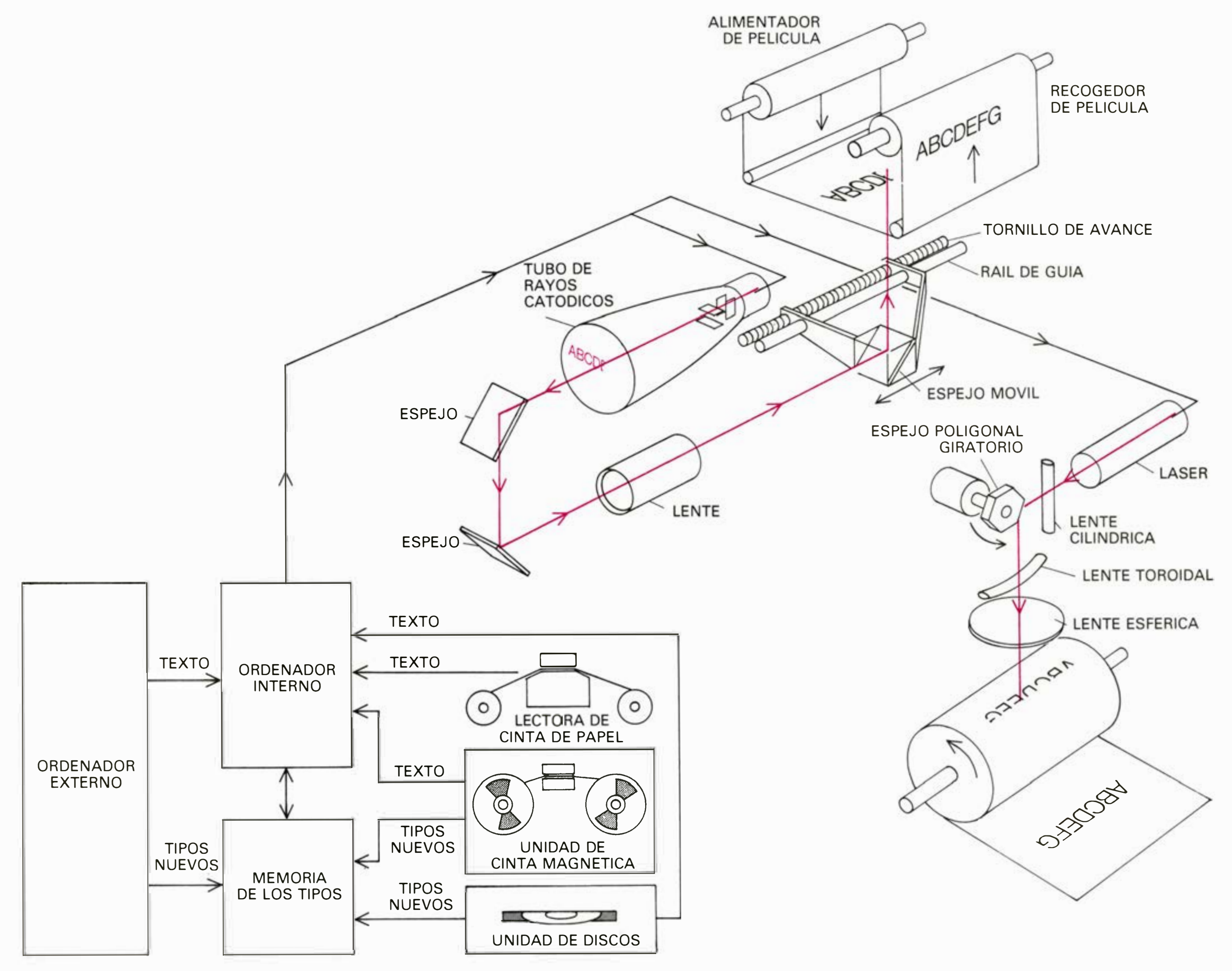
18 PUNTOS

RQEN baegnov

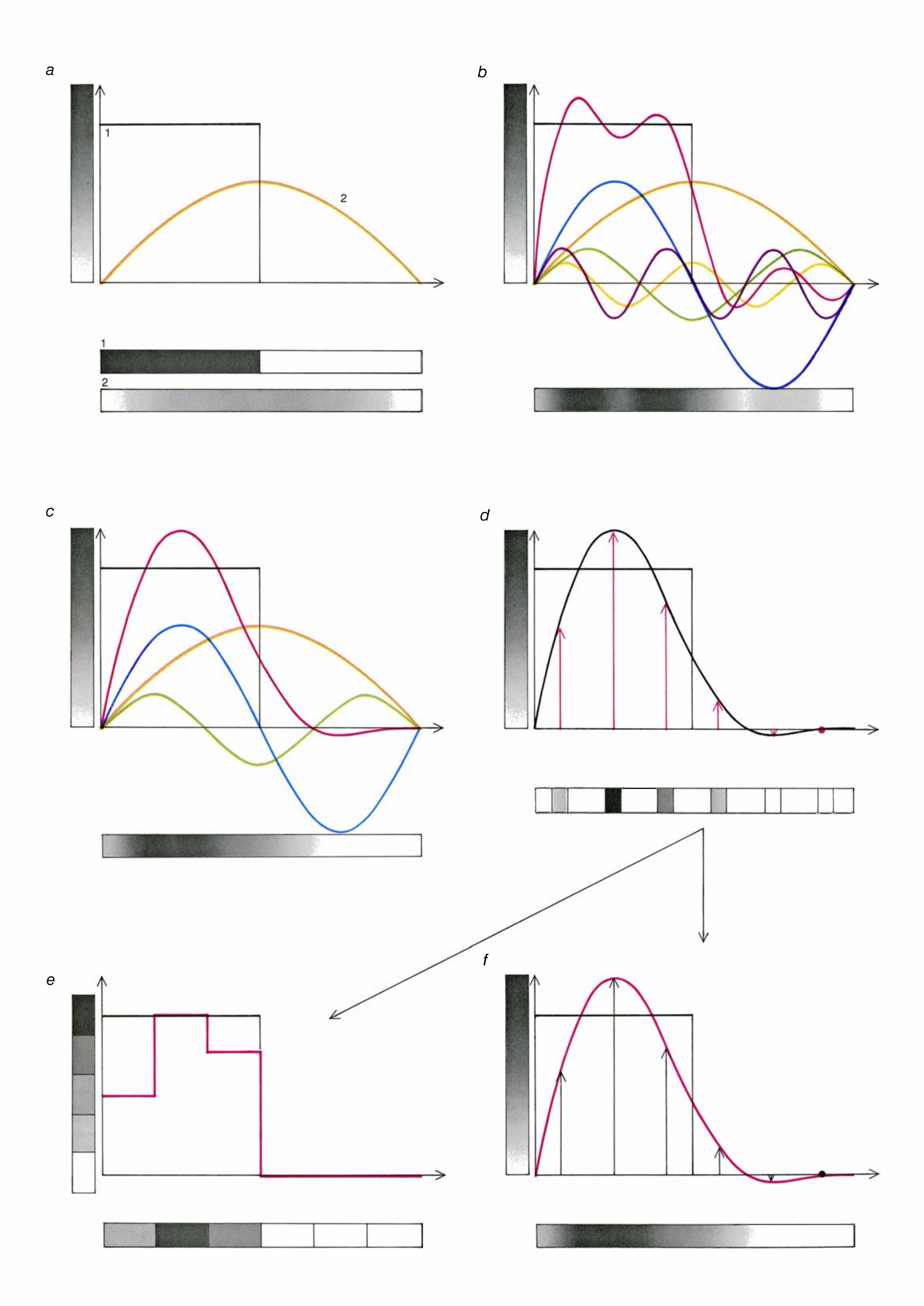
5. LA REDUCCION O LA AMPLIACION A ESCALA del tamaño de los tipos no se puede llevar a cabo de forma estrictamente proporcional, ni siquiera con un solo tipo, si se desea mantener su legibilidad óptima. En los tamaños más pequeños, entre seis y doce puntos, las letras deben ser más anchas y sus trazos largos, trazos finos y remates deben ser más gruesos que en los de sus homólogas de mayor tamaño. En los tamaños más pequeños, trazos tales como el largo de la *b* minúscula y la cola de la *g* minúscula deben ser proporcionalmente más cortos que la altura de la *n* minúscula; el ojo, o espacio interior de las letras, será más abierto y, mayor, el espaciado entre dos letras consecutivas. Los tipos que se reproducen son los denominados tipos romanos "Times"; el tamaño de los tipos se ha variado a escala, mediante el método fotográfico, ajustándolo a la altura de la *n* para facilitar la comparación entre los diseños.

constituyen una sola superfamilia de diseños diferentes de un determinado tipo. Para cada nuevo diseño se especificarán las tres siguientes características: si el tipo es romano o itálico (redonda o cursiva), si su grueso es normal o es negrita y si tienen remates o no los tienen. (Estos remates son los trazos cortos con los que finalizan los trazos más largos de las letras, como es el caso de la pequeña línea horizontal que aparece en la parte inferior de la letra T; evidentemente, los tipos sin remate son los que no poseen estos trazos.) Consideradas conjuntamente, estas tres características generan ocho diseños diferentes. Por otra parte, cada tipo alfabético incluye caracteres de 16 tamaños diferentes. En consecuencia, el número total de glifos o mapas de bits individuales que se necesitan para acomodar un solo carácter con una superfamilia mínima de tipos es de 128; el número de glifos precisos para dar cuenta de toda una superfamilia completa, en la que se incluyan las 128 formas de letras diferentes, es igual a 128^2, es decir, más de 16.000.

Aunque la tecnología digital plantea toda una nueva serie de problemas específicos en el diseño tipográfico, éstos no carecen de precedentes; resulta instructivo explorar la estabilidad relativa de las formas de las letras a través de la historia de la tipografía y sus variaciones a tenor de las presiones impuestas por los sucesivos avances tecnológicos. Subsiguientemente a la introducción de una nueva técnica se producen siempre dos etapas en la evolución del diseño de los tipos. En primer lugar, existe un período de imitación en el que las formas de las letras más sobresalientes de la anterior generación tipográfica sirven de modelos para los nuevos diseños. En segundo lugar, a medida que los diseñadores se van familiarizando con el nuevo medio y adquieren una mayor confianza, surgen nuevos diseños innovadores que no son simplemente imitativos, sino que explotan todas las posibilidades y exploran todas las limitaciones del medio.



6. LAS COMPOSITORAS DIGITALES ALMACENAN un elevado número de tipos para una utilización casi instantánea y pueden producir caracteres de prácticamente cualquier tamaño con una velocidad que alcanza hasta los 15.000 caracteres por segundo. La correspondiente información se incorpora a la máquina mediante una cinta de papel perforado, una cinta magnética, un disco o en forma de un flujo de datos procedentes de un ordenador externo al aparato. En las instalaciones más recientes, el ordenador externo organiza el texto en líneas, columnas y páginas, procede a ajustar el texto, si así se desea, y separa automáticamente las palabras con un guión, siempre que sea necesario, con el fin de obtener un buen espaciado. Una vez especificada la clase de tipo a utilizar, así como el tamaño y la posición de cada uno de los caracteres de una página, los datos pertinentes se convierten en una serie de impulsos electrónicos mediante un ordenador "esclavo", interno, cuya misión consiste en gobernar una máquina de imprimir. En el esquema que se reproduce en esta figura se representan dos clases diferentes de estas máquinas de imprimir. (Como es natural, estos dos dispositivos no se incorporarían conjuntamente en una misma instalación real.) En las compositoras basadas en los tubos de rayos catódicos, un haz de electrones reconstruye cada letra mediante una serie de líneas verticales que se hallan densamente espaciadas. La imagen se proyecta sobre una película o un papel fotosensible. En las impresoras láser, no se utilizan haces de electrones, sino que se recurre a un haz de rayos láser; el láser barre horizontalmente toda una página a la vez y crea una figura de carga eléctrica sobre un tambor. Unas partículas pigmentadas son atraídas por las regiones cargadas e incorporadas al papel mediante calor.


a
1
2
1
2
b
c
d
e
f

Sin embargo, dado que los tipos están destinados en última instancia a ser leídos, la tecnología de la producción de tipos no es el único factor a considerar en la forma final de las letras; el diseño tipográfico es un arte y los mejores diseños son los que reflejan más sutilmente la tensión existente entre la imitación y la innovación.

Las formas de las letras han persistido durante más tiempo que cualquier otro ingenio de uso común. Todavía siguen en uso diseños de letras cuya antigüedad supera los 2000 años; muchos de los tipos más corrientes en la actualidad son réplicas de diseños populares de los siglos XV y XVI. Por ejemplo, los tipos utilizados en la revista *Investigación y Ciencia* fueron diseñados originalmente por los tipógrafos ingleses Stanley Morison y Victor Lardent para el periódico "The Times" de Londres en 1931 y reciben el nombre de Tipos Romanos "Times". Estos diseños estaban basados en ciertos tipos franceses y flamencos introducidos en 1570. Las formas básicas de las letras minúsculas de nuestro alfabeto se crearon durante el siglo VIII en los monasterios y cancillerías del imperio de Carlomagno. Las formas de nuestras letras mayúsculas son sustancialmente las mismas que las que utilizaban los romanos durante el reinado de César Augusto. Más de la mitad de estas mismas letras mayúsculas se han mantenido prácticamente inalteradas desde los tiempos de Pericles, en la Atenas del siglo V antes de Cristo.

En la época de los amanuenses, los perfiles de las letras se creaban mediante una secuencia continua de movimientos de un pincel o de una pluma que se desplazaba sobre la superficie de escritura. Durante dicha secuencia de movimientos, la herramienta utilizada para la escritura ejercía una presión sobre la superficie en cuestión o bien se elevaba sobre ella, generándose así los diferentes trazos que constituyen una letra. Este conjunto de movimientos secuenciales se conoce con el nombre de ductus y define una estructura topológica característica para cada letra del alfabeto. La evolución de la escritura ha estado motivada por cambios en el ductus y por variaciones en la forma y flexibilidad de la herramienta utilizada en la escritura. Esta herramienta es la responsable directa del contraste existente entre los trazos gruesos y finos. Si dicha herramienta es una pluma puntiaguda rígida, el contraste de las letras viene determinado por el ángulo que forma la pluma con la dirección en que se mueve a lo largo de la superficie de escritura. Cuando la herramienta es un pincel o una pluma puntiaguda flexible, el contraste de las letras viene determinado por las variaciones en la presión con que dicho pincel o pluma va describiendo el correspondiente ductus.

Una ventaja importante de la escritura manual es el establecimiento de una relación inmediata entre la forma final de una letra y la persona que la ha escrito. Un buen escribiente es experto en la introducción de ciertos cambios menores en la forma de cada una de las letras con el fin de armonizar una letra determinada con las letras adyacentes a ella. Más aún, para ajustar las líneas de un cierto texto, el escribiente puede insertar determinadas formas de letras o de palabras alternativas o abreviadas. La principal desventaja de la escritura manual radica en que el diseño de cada letra se debe realizar en "tiempo real": cada letra a leer debe escribirse previamente garabateándola a mano.

Cuando se inventaron los tipos móviles en Europa, se crearon unos modelos que reproducían las formas corrientes de la escritura manual en el siglo XV, a pesar de que la creación de dichos tipos utilizaba una tecnología radicalmente diferente. En primer lugar, una imagen de la forma de la letra se grababa en relieve en el extremo de un punzón de acero. A continuación, se golpeaba con este punzón sobre una plancha de cobre con el fin de obtener una matriz en hueco del original. Esta matriz se colocaba en un molde ajustable. Más tarde, sobre el molde y la matriz se vertía una aleación fundida de plomo, estaño, antimonio y cobre. Al endurecerse esta aleación reaparecía el diseño original en relieve. Finalmente, el tipo obtenido se impregnaba con tinta y se presionaba directamente sobre papel. En consecuencia, las letras se creaban mediante un proceso gliptal o escultural. El diseño de la letra se liberaba así de las limitaciones que imponía su ejecución en tiempo real, pero estaba restringido a la necesidad de situarlo sobre una base rígida de forma rectangular. Para imitar las variaciones que sufren las formas de las letras en la escritura manual cuando esas letras se combinan de formas diferentes, algunos de los primitivos tipógrafos fabricaron centenares de variantes diferentes; la producción y composición de estas formas era tan cara que las ventajas económicas que ofrecían los tipos móviles se perdían casi por completo.

La evolución de los perfiles de los tipos gliptales estaba condicionada por ciertas razones perceptuales y conceptuales y no por las necesidades de una escritura rápida. Se prestaba gran atención a la forma de los espacios dentro de las letras y entre ellas. La proporción, la anchura, el grosor y la construcción de una letra se alteraban independientemente de su topología básica, en vez de resultar parcialmente determinados por ella, como sucedía en las letras ductales. Los diseñadores de letras gliptales estaban obligados a hacer una grabación diferente para cada tamaño de letra y, por tanto, al igual que sucedía con sus predecesores amanuenses, podían introducir pequeños ajustes en el espaciado y en la anchura del trazo para los diferentes tamaños de tipos con el fin de facilitar su lectura. Las formas alternativas de las letras se eliminaron casi por completo y las demandas especiales de la nueva tecnología se cubrieron mediante unos diseños de letras que resultaban adecuados en casi todas las combinaciones posibles. Sin embargo, los vestigios de la primitiva técnica ductal se pueden apreciar todavía en los tipos de fundición utilizados a mediados del siglo XX; ciertas combinaciones de letras tales como fi, fl, ffi y ffl se fabricaban en un cuerpo metálico único

7. ANALISIS ESPACIAL DE FRECUENCIAS, susceptible de realizarse en cualquier imagen bidimensional, la de una letra por ejemplo, de la misma forma que una gráfica de casi cualquier función matemática puede aproximarse mediante una suma de curvas seno y coseno. Podemos representar, por ejemplo, un rectángulo de bordes perfectamente definidos a través de una línea recta paralela al eje horizontal de la gráfica; la densidad de gris o de negro en el rectángulo corresponde a la altura de la línea sobre el eje horizontal. Para aproximar el rectángulo se puede superponer un conjunto único de ondas seno y coseno. La primera onda seno de la aproximación se representa en la figura *a*. La onda espacial, variación desde el blanco al negro a través de diferentes tonalidades de gris y retorno de nuevo al blanco, que corresponde a una onda seno, es una aproximación bastante tosca de dicho rectángulo. Sin embargo, conforme aumenta el número de componentes seno y coseno la aproximación es mejor: cuando se suman las cinco ondas seno coloreadas de diferentes alturas y longitudes de onda se obtiene la curva suave de color rojo que se representa en la figura *b*. Para eliminar el ruido de alta frecuencia, las formas resultantes se pueden filtrar electrónicamente suprimiendo las componentes por encima de una determinada frecuencia. En la figura *c* se han eliminado las dos componentes de frecuencias más altas de la curva en rojo representada en *b*; es decir, las ondas seno de color amarillo y púrpura, obteniéndose la curva roja de *c*. La curva filtrada se muestrea determinando su altura, o valor en una escala de grises, para una serie de intervalos regularmente espaciados a lo largo del eje horizontal (*d*). Asignando un valor discreto de gris a cada intervalo se obtiene una aproximación digital a la curva filtrada (*e*). Se puede demostrar matemáticamente que la curva filtrada original es reconstruible a partir de las muestras originales por interpolación entre los correspondientes puntos de muestreo, siempre y cuando la frecuencia para la que se toman las muestras sea mayor que el doble de la componente de frecuencia más alta en la curva filtrada (*f*).

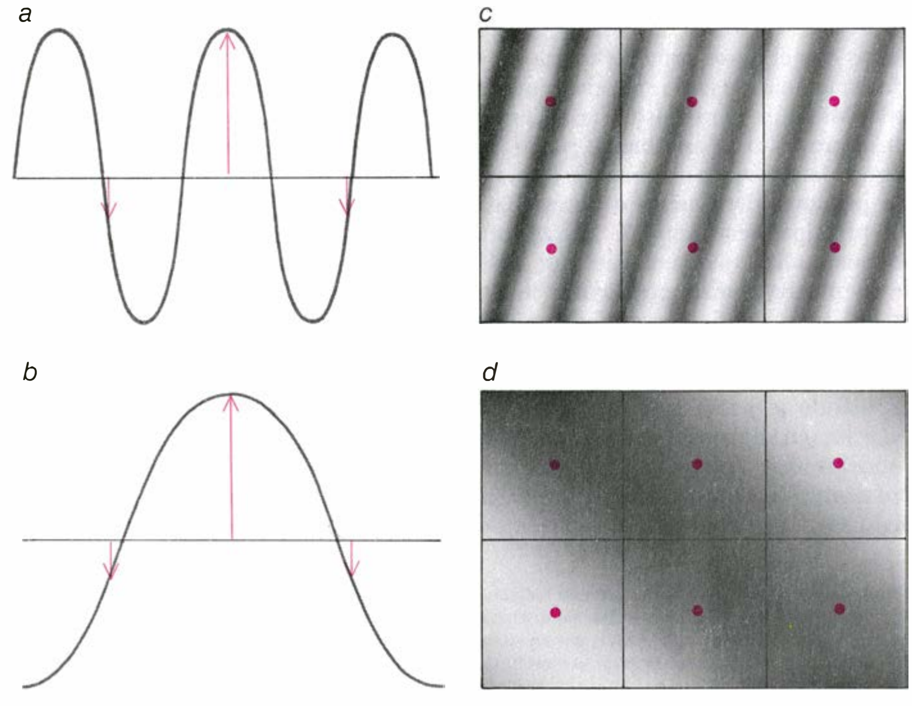
o politipo. Para conseguir un buen espaciado en combinaciones tales como TA y AY, el cuerpo de los tipos utilizados solía ajustarse a mano, lo cual representaba un coste considerable.

Con la llegada de la fotocomposición a mediados de la década de 1950, la tipografía se enfrentó una vez más con un nuevo conjunto de variables tecnológicas. En los primitivos sistemas de fotocomposición los destellos de un foco de luz estroboscópico incidían sobre el negativo de una letra patrón grabada sobre una película o vidrio transparente y posteriormente atravesaban una lente que proyectaba una imagen de la letra patrón sobre un papel o película fotosensibles. El tamaño de la imagen de la letra se controlaba intercambiando varias lentes o utilizando un ampliador óptico, o zoom, de distancia focal variable. Las placas de impresión se obtenían al revelar la imagen que se había formado en el papel o en la película. Dado que la posición de la letra sobre el papel se podía controlar mediante una lente y un prisma o espejo que se movía a pequeños pasos a lo largo de una línea, el ajuste del espaciado entre las letras para ciertas combinaciones especiales era mucho más fácil que en el caso de los tipos metálicos. Por otra parte, la introducción de esta técnica supuso la pérdida de una ventaja significativa de la tipografía por metal fundido. El tipo de diseño no se podía variar con el tamaño de la letra sin preparar una nueva imagen patrón. El coste de los portadores adicionales de imágenes patrón y las dificultades que planteaban los oportunos intercambios limitaron la utilización de esta técnica.

Los problemas a los que se enfrentaba el tipógrafo que tenía que trabajar sometido a las restricciones de todas estas técnicas eran muy similares a los que se le plantean en la actualidad al diseñador de tipos digitales. Por ejemplo, cualquier versión tipográfica de los tipos de letras que son más familiares para los lectores de hoy en día será un sistema de comunicación adecuado siempre que logre conjugar dos características opuestas: la discriminabilidad y la similaridad de los caracteres alfabéticos. Para evitar las confusiones, el lector de una letra debe ser capaz de distinguirla con rapidez y sin ambigüedad de cualquier otra. Por otro lado, las formas de las letras deben poseer también toda una serie de rasgos gráficos comunes: si una letra fuera demasiado distinta de las que constituyen un determinado tipo, el curso de la lectura resultaría perturbado. Donald E. Knuth, de la Universidad de Stanford, ha expresado muy sucintamente los objetivos fundamentales del diseño de letras: "Una letra de imprenta debe ser sublime en su apariencia y subliminal en sus efectos".

A lo largo de la historia de la escritura, las exigencias económicas han entrado siempre en contradicción con las exigencias en producir una tipografía más bella y más fácil de leer. Como siempre, la necesidad debe adecuarse a los propósitos. En el caso de ciertos documentos, se sacrificará la calidad de las letras en aras de una reproducción más rápida, más económica y más compacta. Por ejemplo, la factura de una venta puede abaratarse en proporción y almacenaje, pues se leerá poco y por poca gente; la producción lenta y cara de una inscripción está justificada si esa inscripción está destinada a ser leída con frecuencia y por mucha gente. En tiempo de los romanos, el alfabeto básico se escribía rápidamente utilizando una letra cursiva semilegible cuando se utilizaba en un documento de papiro; en cambio, en los monumentos se utilizaban letras mayúsculas de tipo imperial grabadas a cincel. En la actualidad, los impresos baratos por matrices de puntos resultan menos legibles que la tipografía que se utiliza en los anuncios publicitarios y que resulta mucho más cara.

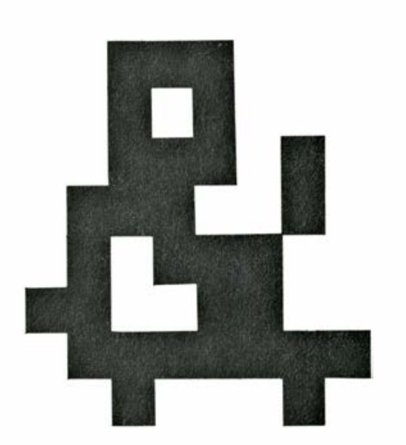
Los efectos que las variaciones complejas en el diseño de las letras producen sobre los lectores no han sido todavía cuantificados; por el contrario, la respuesta del sistema visual a las variaciones espaciales de la intensidad luminosa se ha estudiado durante más de dos décadas. La variación más sencilla de analizar es la analogía visual de un tono acústicamente puro, cuya variación periódica de la intensidad con el tiempo se puede representar en forma de una onda seno o coseno. Un tren de ondas seno o coseno se puede visualizar recurriendo a una cinta comprimida a lo largo de su longitud, de forma que sus bordes representen una serie de ondas corrientes de tipo seno o coseno. Si el lado superior de la cinta se oscureciera de forma tal que las crestas fueran totalmente negras, el lado inferior se mantuviera blanco y las zonas interme-

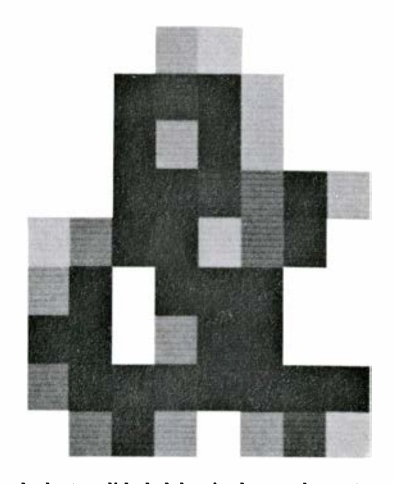


8. ESTA FIGURA PRETENDE ILUSTRAR LA FORMA EN QUE SE PRODUCE EL RUIDO, o señal no deseada, más frecuente en la tipografía digital. Cuando se muestrea la altura de una onda, es decir, los valores de la escala de grises de una imagen, se pierde una parte de la información que contiene la curva o la imagen original. La reconstrucción de la onda o la imagen a partir de unos determinados puntos de muestreo exige que los valores que corresponderían a los puntos situados entre dichos puntos de muestreo se obtengan por interpolación, calculando las componentes seno y coseno cuya superposición reproduzca los valores de los puntos muestreados. Sin embargo, si la onda original se muestrea con una frecuencia inferior al doble de la frecuencia de su componente de mayor frecuencia, las componentes de alta frecuencia que se obtengan en la reconstrucción se corresponderán con otra frecuencia, inferior a la componente de frecuencia más alta de la imagen original. En otras palabras, componentes espúreas de baja frecuencia substituirán las frecuencias auténticas. Por ejemplo, si una onda seno se muestrea sólo 1,5 veces por ciclo (*a*), una reconstrucción que ajuste los valores de las muestras (*flechas coloreadas*) es una onda seno cuya frecuencia es la mitad de la original (*b*). En una onda espacial de dos dimensiones, las muestras son los diferentes valores de una escala de grises tomados en el centro de los cuadrados (*c*). La reconstrucción de una imagen a partir de dichos valores origina un ruido, una onda espúrea de frecuencia inferior a la de las ondas originales y cuya orientación resulta también diferente (*d*). Como es de suponer, la cuantización de las imágenes de las letras se verá afectada por los errores que indroduce dicho ruido.

dias representaran las diferentes tonalidades de gris, las bandas borrosas paralelas alternativamente claras y oscuras que se observarían desde la parte superior de la cinta formarían un tren de ondas sinusoidales en tres dimensiones.

Un tono musical puro se caracteriza por su amplitud o intensidad y por su frecuencia o tono. La amplitud de una onda espacial seno o coseno representa el contraste máximo, o desviación respecto al gris neutro, que se pone de manifiesto entre las partes más claras y más oscuras del tren de ondas, en tanto que la frecuencia es el número de variaciones de luz a oscuridad y de nuevo luz que se producen en una longitud dada. Los psicofísicos han medido la capacidad del sistema visual para distinguir bandas sinusoidales de diferentes contrastes y frecuencias de un campo uniformemente gris. Estos investigadores han encontrado que la sensibilidad a la variación espacial de la luz y la oscuridad depende de la frecuencia de dicha variación: la sensibilidad es máxima cuando la frecuencia espacial es aproximadamente igual a tres ciclos por grado de ángulo visual; y no se puede percibir ningún contraste, por fuerte que sea, bajo la mayoría de condiciones en que la frecuencia es mayor que 60 ciclos por grado. (La detección de cables telefónicos destacándose sobre el cielo es uno de los casos relativamente raros en que se pueden discernir frecuencias espaciales que son quizá mayores de 60 ciclos por grado.)





9. LOS ERRORES DE CUANTIZACION se producen cuando la tonalidad del gris de un elemento gráfico discreto mínimo, pixel, no se ajusta al tono de gris que existe realmente en el punto de muestreo que corresponde a ese elemento gráfico concreto en la figura orginal de una letra. Por ejemplo, los elementos gráficos utilizados en la tipografía digital habitual deben ser blancos o negros; este hecho se basa en que los bordes de las letras originales están aparentemente bien definidos. En realidad, las frecuencias espaciales que son demasiado elevadas para ser percibidas por el sistema visual deberían ser filtradas antes de que las letras sean muestreadas y digitalizadas. Los bordes de las letras filtradas ya no son nítidos; por el contrario, en estas zonas se pone de manifiesto la existencia de tonos grises que sirven de transición continua entre el blanco y el negro. Cuando las letras se reproducen en una máquina de baja resolución, tal como la terminal de un tubo de rayos catódicos, la degradación aparente de la letra producida por el muestreo de baja resolución se puede reducir si los elementos gráficos reproducen con exactitud la tonalidad de gris existente en cada punto de muestreo. Los elementos gráficos que constituyen la letra que se reproduce a la izquierda de la figura son blancos o negros y el diseño resulta casi ilegible. Los elementos gráficos (pixels, en el argot tipográfico) que constituyen esa misma letra en el diseño de la derecha pueden tomar uno de 16 valores posibles de gris. Bizqueando o mirando el diseño de la derecha a una distancia de unos ocho metros se puede percibir claramente la forma de un signo &. Los diseños digitales utilizados en esta figura han sido preparados por John E. Warnock, de Xerox Corporation.

¿Cuál es la importancia de todos estos descubrimientos para la lectura? Aunque la variación espacial del blanco al negro en las letras no es del tipo sinusoidal, las figuras rítmicas fundamentales son evidentes en las letras. Por ejemplo, cuando la letra minúscula *n* se barre de izquierda a derecha a la altura de su centro, el brillo de la imagen varía suavemente de la luz a la oscuridad al llegar al primer trazo vertical, de nuevo luz en el ojo, o interior, de la letra, otra vez oscuridad en el segundo trazo vertical y, finalmente, otra vez luz cuando se llega a la derecha de la letra. Dado que la lectura se hace frecuentemente bajo condiciones de iluminación bastante precarias, es natural que la frecuencia fundamental de las letras de texto haya variado, de tal forma que la sensibilidad del sistema visual alcance su máximo valor.

Este ajuste es prácticamente el que se encuentra en los textos impresos. Cuando se lee un escrito, la imagen sobre la que recae la máxima atención se proyecta en la fóvea, es decir, en la zona más sensible de la retina. La fóvea forma un ángulo de uno o dos grados y, a una distancia de unos 30 centímetros, el ángulo que define equivale a una distancia lineal sólo ligeramente mayor que la longitud de una palabra de cinco letras si se utilizan tipos de 10 puntos, es decir, tipos del tamaño más frecuente. Dado que, por término medio, hay unos 10 ciclos en una palabra de cinco letras, la frecuencia espacial de la imagen que llega a la fóvea es de unos cinco o diez ciclos por grado, un valor que resulta algo superior a la frecuencia de máximo contraste del sistema visual. Los datos existentes parecen apoyar la conclusión de que en una lectura rápida grupos de palabras con más de cinco letras se pueden leer con una sola fijación del ojo. En este caso, la correspondiente imagen se proyecta parcialmente sobre la parafóvea (región de la retina que rodea a la fóvea).

Hasta aquí hemos supuesto tácitamente que las figuras de blanco y negro que forman las letras son ondas espaciales sinusoidales, pero tal vez no resulte tan evidente que el modelo de onda sinusoidal soporte un análisis más detallado de esta cuestión. Sin embargo, al igual que sucede con los sonidos complejos que produce una orquesta sinfónica, susceptibles de ser analizados como una suma de armónicos (tonos puros de frecuencias e intensidades diferentes), se puede analizar prácticamente cualquier forma como una combinación de muchas ondas espaciales seno y coseno. Este resultado se basa en un teorema del matemático francés Jean-Baptiste Joseph Fourier. En 1807, Fourier demostró que, superponiendo una serie de ondas seno y coseno unidimensionales de diferentes fases, amplitudes y frecuencias, la gráfica de casi cualquier tipo de función se puede aproximar hasta el límite de precisión que se desee.

Las componentes sinusoidales espaciales en dos dimensiones poseen un grado adicional de libertad: su orientación respecto a una dirección fija en el plano. El teorema de Fourier demuestra que se puede reconstruir casi cualquier figura (por ejemplo, una letra) con cuanta precisión se desee superponiendo una serie de ondas seno y coseno con una fase, amplitud, frecuencia y orientación adecuadas. En general, las elevadas frecuencias espaciales que conlleva la imagen de una letra están originadas por los bordes de las letras y por ciertos detalles menores como los remates y el estrechamiento gradual de los trazos principales de la letra. Como ya hemos indicado antes, las frecuencias espaciales más bajas definen el perfil de una letra, es decir, su aspecto general, en blanco y negro. El espectro de las frecuencias espaciales que se necesita para representar una letra determinada se conoce como la anchura de

banda de esa letra. Si nos basáramos exclusivamente en el dato de que la frecuencia máxima para la que el ojo humano es capaz de apreciar un contraste es de 60 ciclos por grado de ángulo visual, llegaríamos a la conclusión de que la obtención de una reproducción tipográfica perfecta exigiría la eliminación de todas las frecuencias espaciales de una letra que fueran superiores a los 60 ciclos por grado.

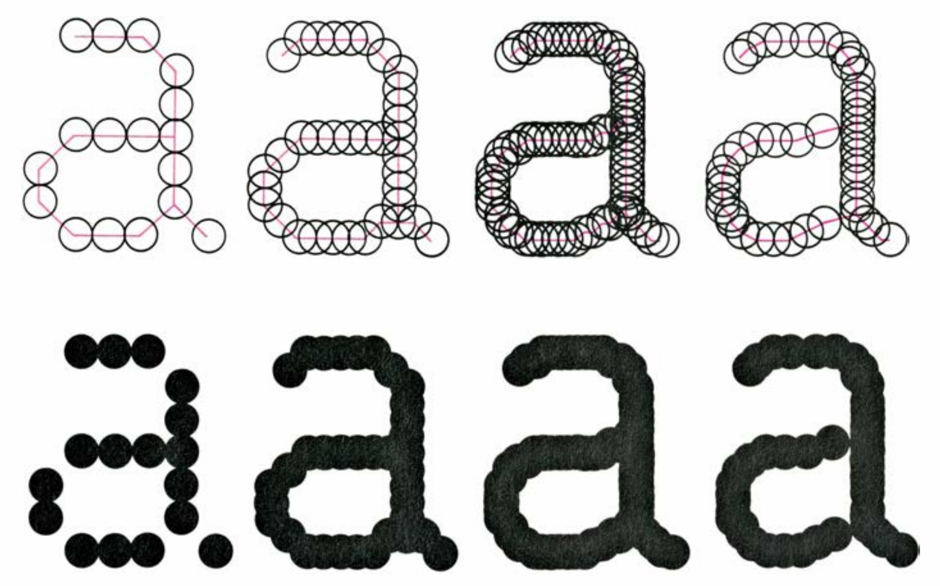
En la práctica, la tipografía digital de alta calidad exige que las letras incluyan frecuencias espaciales elevadas, de 120 ciclos por grado, como mínimo. Una explicación se encuentra en el hecho de que para ciertos detalles finos, tales como las terminaciones de los remates, la agudeza del sistema visual puede exceder los 60 ciclos por grado. Sin embargo, la principal razón es la relación entre las frecuencias espaciales que constituyen una determinada letra y el proceso de digitalización de esa letra. Para digitalizar una letra es necesario reducirla a un conjunto de unidades discretas. Por ejemplo, si los bordes de la letra fueran perfectamente nítidos, la letra se podría muestrear superponiéndole una rejilla de cuadrados y observando si el punto central de cada uno de los cuadrados coincide o no con algún punto de la letra. Si el punto central de un cuadrado coincide con un punto de la letra, todo el cuadrado se ennegrece y si el punto central no coincide con ningún punto de la letra, el cuadrado se deja en blanco.

Un borde perfectamente nítido exigiría un número infinito de senos y cosenos para representar toda su anchura de banda espectral. Por otro lado, si la anchura de banda de una letra se limita de forma tal que no incluya las componentes sinusoidales por encima de una cierta frecuencia, sus bordes aparecerán ligeramente borrosos. La superposición de un número finito de componentes sinusoidales proporciona una transición suave desde el negro hasta el blanco a través de distintas tonalidades de gris. Sin embargo, en las letras con banda limitada se obtiene una ventaja de índole matemática: de acuerdo con los resultados encontrados por Harry Nyquist en los Bell Telephone Laboratories en 1924 y 1928, y desarrollados por Claude E. Shannon en el mismo centro, en el año 1949, si se conoce la componente de frecuencia más alta de una señal de banda limitada, y si las muestras se toman con igual espaciado y con una frecuencia mayor que el doble de la frecuencia componente más alta, las letras con banda limitada pueden reproducirse del todo a partir sólo del correspondiente muestreo de puntos. Por tanto, teóricamente, una letra cuya frecuencia más elevada sea de 60 ciclos por grado de ángulo visual, se reconstruirá totalmente a partir de los diferentes valores de gris existentes en puntos regularmente espaciados muestreados con una frecuencia ligeramente mayor que 120 veces por grado de ángulo visual.

Una de las dificultades que presenta la frecuencia teórica de muestreo radica en el error que introducen los instrumentos que examinan y muestrean las letras. Los puntos de muestreo pueden no estar regularmente espaciados y los valores medidos de gris pueden no ser suficientemente precisos. En la práctica, la imposibilidad de una limitación brusca de la banda plantea un problema todavía más serio. De todo ello se desprende que en la tipografía de alta calidad el muestreo se lleva a cabo con una frecuencia no inferior a 240 por grado de ángulo visual, suficiente en principio para reconstruir una letra con banda limitada de 120 ciclos por grado. Este ritmo de muestreo equivale a una resolución de 200 líneas por eme para un tipo de 12 puntos, es decir, unas 1200 líneas por pulgada.

Si la imagen de una letra se muestrea de forma insuficiente, es decir, a menos del doble de la frecuencia de su componente sinusoidal más alta, los valores de la escala de gris de las muestras pueden ser idénticos a los valores que resultarían de una frecuencia espacial más baja. Cuando se trabaja en estas condiciones, se origina un ruido y se pierde inevitablemente información. Las ondas espaciales de baja frecuencia que no figuran en el espectro de ondas de la letra reemplazan a las componentes originales de alta frecuencia; una reconstrucción de la letra inicial a partir de tales muestras incorporaría también las componentes espúreas de baja frecuencia. Este fenómeno constituye la fuente más importante de ruido o distorsión en la tipografía digital. Dicho fenómeno se pone de manifiesto con absoluta claridad cuando una letra se convierte en una fila de elementos gráficos como resultado del muestreo; su importancia depende de la tosquedad de tal muestreo.

Cuando se trabaja a alta resolución, el fenómeno al que acabamos de referirnos se manifiesta en forma de una ligera rugosidad en los contornos de la letra; para una resolución media, los contornos aparecen recortados de forma desigual, y a baja resolución las curvas se hacen poligonales y las diagonales se dislocan. Para resoluciones todavía más bajas, las diferencias entre los elementos rectos, curvos y diagonales desaparecen y las letras resultan prácticamente ilegibles. Más aún, a medida que la resolución disminuye, se reduce la diversidad de los posibles dise-



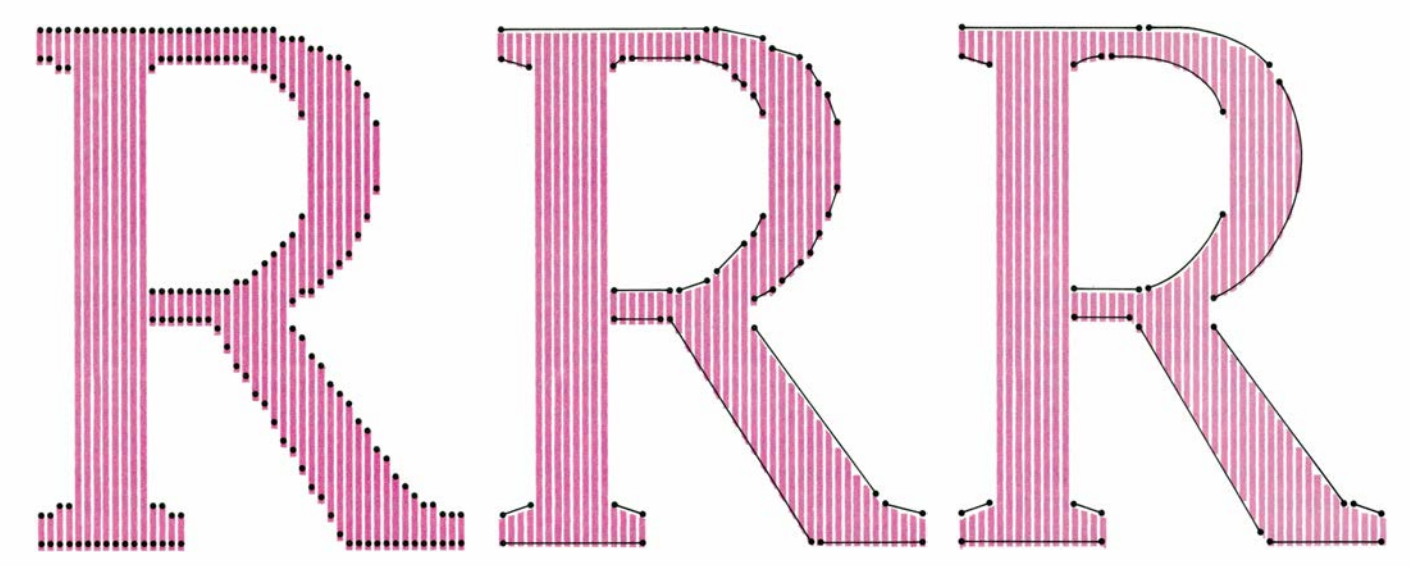
10. EL TAMAÑO DEL PUNTO DE ESCRITURA de una impresora digital, tal como una impresora de matriz de puntos, determina el tamaño del elemento más pequeño que se puede diferenciar en una letra, aunque la posición del punto en cuestión todavía se puede ajustar para así mejorar la imagen. Las filas de puntos paralelos que constituyen la cabeza impresora se pueden inclinar o hacer que la cabeza recorra varias veces un mismo carácter. A la izquierda de esta figura se ha reproducido la letra *a* minúscula, sin que se produzca solapamiento alguno en los puntos de escritura; el resultado es una letra de baja calidad. En la segunda letra, a partir de la izquierda, el solapamiento de dichos puntos ha aumentado en un 50 por ciento y en la tercera letra hasta un 75 por ciento; las letras resultantes se manifiestan más continuas y los bordes de las mismas están mejor definidos. Sin embargo, a medida que el solapamiento aumenta, las líneas horizontales se hacen demasiado gruesas en relación con las líneas verticales y las uniones resultan excesivamente ennegrecidas. En la parte derecha de la ilustración se han eliminado algunos de los puntos con el fin de aligerar un tanto las líneas horizontales y afinar las uniones. (Dibujos de Allen Beechel.)

ños de letras; para resoluciones muy bajas, tales como las matrices de puntos de cinco por siete o de seis por nueve, que se suelen utilizar en muchas terminales de tubos de rayos catódicos y en impresoras por matrices de puntos, el número de variantes posibles es muy bajo.

Una consecuencia del muestreo de baja resolución menos evidente, pero en última instancia más importante, es la pérdida de información acerca de las relaciones de proporcionalidad que se reflejan en el diseño de un alfabeto. En un diseño de alta calidad, adecuadamente ajustado a las características del sistema visual, existe una sutil interrelación entre la proporción y el grosor de los diferentes trazos de las letras, así como entre sus elementos gruesos y delgados. La relación entre la anchura de trazo y la altura de la letra *x* minúscula y también la anchura media de las letras se ajusta con todo cuidado. El resultado de todos estos detalles de diseño es difícilmente perceptible en una letra individual, pero al contemplar un texto extenso la calidad de un diseño se pone de manifiesto en forma de una armonía visual entre las letras negras y los espacios blancos, así como en la existencia de un nivel de gris muy agradable en el conjunto del texto. Con una alta resolución digital todas estas variaciones proporcionales se reproducen de forma muy precisa y la textura tipográfica parece sólo un poco menos refinada que la de una tipografía analógica. Sin embargo, a medida que la resolución digital de una letra disminuye, las variaciones proporcionales se igualan, los elementos tipográficos se homogeinizan y el aspecto del texto resultante se hace más tosco y desagradable.

Cuando el coste económico o las limitaciones técnicas del dispositivo de salida exigen que las letras se almacenen o reproduzcan con una resolución inferior a la ideal, quedan todavía diferentes estrategias para mejorar la calidad de la imagen de las letras. Por ejemplo, si los tonos de gris medidos en cada punto de la muestra se mantienen a lo largo de todo el elemento gráfico (pixel) que corresponde a ese punto, las líneas con un recorte desigual se hacen mucho menos aparentes en la letra final de lo que lo serían si el elemento gráfico fuera blanco o negro. Ciertos dispositivos con baja resolución espacial son capaces de reproducir los elementos gráficos con diferentes tonos de gris y, por consiguiente, el denominado error de matización entre la tonalidad de un pixel y la del punto de muestreo puede resultar bastante pequeño. Por ejemplo, la pantalla terminal de un vídeo puede tener 256 o más niveles de gris en cada elemento gráfico; este hecho ayuda a compensar en parte la baja resolución del muestreo.

Una segunda estrategia para mejorar la imagen de una letra consiste en almacenar la letra en cuestión en la memoria de un ordenador en una forma tal que se aproveche la regularidad general de su diseño. Por ejemplo, la mayoría de las letras consisten en la conexión de ciertas formas simples; la información sobre los valores de la escala de grises



11. ESPACIO DE ALMACENAMIENTO en la memoria de un ordenador necesario para reproducir una letra con una determinada resolución; puede reducirse de diferentes maneras. Una de las posibles estrategias se basa en almacenar tan sólo las posiciones del final de cada uno de los trazos de un haz de electrones o de un láser, en lugar del valor resultante de una escala de grises para cada elemento gráfico (*izquierda*). La codificación de trazos permite reducir todavía más las exigencias de memoria: el perfil de una letra se puede especificar considerando ciertos puntos críticos conocidos como nudos de unión de los trazos. Cuando se necesita utilizar esa letra, un ordenador lleva a cabo la interpolación entre líneas rectas (*centro*) o entre líneas rectas y arcos de circunferencia (*derecha*) existentes entre los nudos de unión de los trazos con el fin de generar los puntos finales de cada trazo en el instrumento de escritura. Con el fin de obtener una aproximación todavía mejor a la forma de las letras originales, se pueden interpolar otras curvas más elaboradas: las espirales logarítmicas; pero su uso se traduce en una reducción en la memoria del ordenador disponible para efectuar cálculos subsiguientes.

de las muestras tomadas en el interior de uno de los trazos de una letra es redundante debido a que todas las muestras son negras. De aquí que la capacidad de la memoria de un ordenador que se necesitaría para almacenar una escala arbitrariamente compleja de valores de gris para los elementos gráficos que corresponden a un trazo se puede destinar a almacenar más información sobre sus límites. Las letras se pueden codificar también estructuralmente en forma de diferentes combinaciones de las porciones que corresponden a los trazos de una pluma o pincel en las letras escritas a mano. Una codificación más compleja de la regularidad en las letras requiere menos memoria del ordenador; en cambio, la correspondiente decodificación posterior es más laboriosa.

Una de las formas más versátiles para reducir el número de muestras a almacenar se basa en registrar sólo las posiciones de algunos puntos seleccionados a lo largo del perfil de la letra. Estos puntos se pueden unir construyendo matemáticamente unas líneas rectas o curvas cuyas formas se calculan a partir de los puntos básicos cuando la letra se recupera de la memoria de un ordenador. Si sólo se utilizan líneas rectas, las curvas de las letras se convierten en polígonos; en tal caso, los segmentos rectos deben ser lo suficientemente cortos para que no se perciba la poligonización. Las líneas de unión curvas que se deducen de las gráficas de ecuaciones cuadráticas, tales como la circunferencia, la elipse o la parábola, proporcionan una mejor aproximación a los perfiles complejos dibujadas por la mano que la que proporcionan las líneas rectas. Las líneas curvas de orden superior, así las que se obtienen a partir de las gráficas de ecuaciones cúbicas, pueden proporcionar aproximaciones todavía mejores, pero la complejidad del cálculo necesario para construir las correspondientes curvas es también mayor.

Cuando se almacena una letra bajo la forma de un conjunto de puntos, o nudos de líneas, las coordenadas de esos nudos se pueden codificar con la ayuda de un dispositivo llamado digitalizador, con resoluciones de hasta 8000 líneas por eme, mucho mayor que la de cualquier dispositivo de salida de textos. Los perfiles de los caracteres resultantes se pueden usar entonces como patrón para generar una matriz de elementos gráficos susceptibles de ser utilizados en cualquier clase de barrido. Dicho conjunto de nudos puede servir también de forma patrón de una letra a partir de la cual un ordenador generará una gran variedad de versiones de dicha letra. Por ejemplo, las líneas básicas de orden superior se pueden convertir en arcos circulares o en líneas rectas con la ayuda de ciertos dispositivos capaces de generar el perfil de una letra en vez de un mapa de bits. Los contornos de las letras se pueden modificar de escala en un sentido creciente o decreciente y ensanchar en cualquier dirección. Si la posición de cada nudo se complementa con una etiqueta descriptiva que establezca su papel en la forma de la letra considerada en relación con los restantes nudos, un cierto tipo se puede ajustar automáticamente sobre una plantilla determinada sin necesidad de llevar a cabo un trabajo tan tedioso como el de ir encendiendo o apagando los elementos gráficos (pixels) a mano.

Se han escrito varios programas de ordenador para manipular la forma de las letras basadas en las líneas básicas. El sistema Ikarus, desarrollado por Peter Karow, de la URW Unternehmensberatung de Hamburgo, se utiliza ampliamente en la industria tipográfica. En una rejilla electrónica se coloca un dibujo preciso del perfil de una letra sobre la cual se han señalado los nudos; las posiciones de los nudos se introducen en la memoria de un ordenador mediante un cursor. Los diferentes elementos que constituyen las letras, como los remates, se identifican a medida que se van introduciendo en la memoria. El programa Ikarus calcula las líneas básicas y realiza automáticamente toda una serie de variaciones en el diseño: cambio en el tamaño de una letra en una determinada figura de barrido, modificación del grosor de los trazos de una letra desde los más finos hasta los más gruesos o interpolación entre dos formas diferentes de una misma letra.

Donald Knuth ha desarrollado otro sistema de diseño basado en las líneas

fundamentales de las letras y conocido con el nombre de Metafont. El lenguaje utilizado para programar el Metafont está basado en los principios ductales; una vez que se ha descrito la topología de una letra, su forma final se determina especificando las características de una "pluma" virtual que recorre el esqueleto de la letra en cuestión. (Esta pluma virtual no se debe confundir con ninguna pluma real; el ordenador se limita simplemente a representar la forma final de la letra como si hubiera sido dibujada por una pluma real.) El tamaño de la pluma virtual, el ángulo que forma en relación con la superficie de escritura, las características de su extremo y otras variables susceptibles de ser especificadas independientemente, dan lugar a una serie de contornos de letras bastante diferentes a partir de un mismo esqueleto. Actualmente están en fase de desarrollo otras versiones del programa Metafont que son capaces de describir las formas de las letras de acuerdo tanto con los principios gliptales como con los principios ductales.

Como es natural, se han puesto a punto otros muchos sistemas además de los citados anteriormente. El conocido con el nombre de FRED, un programa basado en líneas básicas cúbicas escrito por Patrick Baudelaire, y el denominado Prepress, un programa de edición de elementos gráficos escrito por Robert F. Sproull, se desarrollaron en el Centro de Investigación de Palo Alto de la Xerox Corporation. Un sistema de líneas básicas espirales fue desarrollado por Peter Purdy y Ronald McIntosh de la firma Purdy y McIntosh, de Watford, Inglaterra; el sistema ELF, basado en principios ductales, fue escrito por David Kindersley y Neil Wiseman.

A pesar de su aceptación casi universal y de su ampliamente reconocida flexibilidad, sólo ahora la tipografía digital empieza su transición desde la fase de imitación a la de innovación. De hecho, el muestreo de letras y todo el aparato teórico en que se basa dicho muestreo presuponen ya la existencia de un diseño de letras. Este modelo de letras es de tipo analógico, y el éxito de la tipografía digital se juzga a la luz del grado en que es capaz de imitar las letras ductales o gliptales. Sin embargo, los problemas que plantea el muestreo y la digitalización de las letras parecen sugerir que tal vez sería más productivo diseñar nuevas formas de letras utilizando directamente la tecnología digital. Por otra parte, los avances recientes en el estudio de la visión abren la posibilidad de que el diseñador de tipos digitales experimente las diferentes formas de adaptación de la tipografía a las necesidades reales del sistema visual. Si se pudieran desarrollar imágenes de letras que se aproximaran más estrechamente al "lenguaje" de la percepción visual, se incrementaría sustancialmente la velocidad y la eficiencia de la lectura.

La mejor forma de producir un diseño creativo se basa en un sistema sintético, en el que el diseñador de letras puede interaccionar rápidamente con un ordenador con el fin de observar inmediatamente sobre una pantalla los efectos de los cambios introducidos en el diseño, de manera análoga a como un diseñador tradicional observa y corrige sobre la marcha el trazo de una pluma o de un pincel sobre un papel. El sistema debe ser preciso, de alta resolución y capaz de reproducir instantáneamente las múltiples versiones posibles

de un diseño, tales como las líneas básicas del perfil de una letra, los mapas de bits, las matrices de elementos gráficos con diferentes tonalidades de grises o la simulación de la salida de una impresora determinada. Todavía no existe un dispositivo que reúna todas estas características, pero la aparición de una generación de circuitos integrados especiales para las artes gráficas pueden favorecer su construcción en un plazo breve. Cuando se disponga de tales sistemas aparecerán nuevas formas de letras y se iniciará la fase creativa de la era digital, de modo análogo a como sucedió con las eras ductal y gliptal.

a

Hamburgefons Hamburgefons

b

Hamburgefons Hamburgefons

c

a a a a a a

d

A pen of aspect 1/3 generated these letters.
A pen of aspect 2/3 generated these letters.
A pen of aspect 1/1 generated these letters.

e

The x-height and the heights of ascenders and descenders can be independently specified.

f

A 'slant' parameter transforms the pen motion, as shown in this sentence, but the pen shape remains the same. The degree of slant can be negative as well as positive, if unusual effects are desired. *Too much slant leads, of course, to letters that are nearly unreadable.* Perhaps the most interesting use of the slant parameter occurs when Computer Modern Italic fonts are generated without any slant.

g

The 'square root of 2' in these letters is 1.100.
The 'square root of 2' in these letters is 1.300.
The 'square root of 2' in these letters is 1.414.
The 'square root of 2' in these letters is 1.500.
The 'square root of 2' in these letters is 1.700.

12. SE PUEDEN REALIZAR AUTOMATICAMENTE VARIACIONES EN EL DISEÑO de letras digitales con la ayuda de diferentes programas de ordenador. En el sistema Ikarus, desarrollado por Peter Karow, de la URW Unternehmensberatung de Hamburgo, las letras se amplían o reducen sin alterar la anchura de sus trazos (*a*). Este programa permite también corregir los diseños para los diferentes tamaños de impresión; los cambios necesarios son mucho más complejos que la simple dilatación o compresión de la imagen en una determinada dirección (*b*). Se pueden producir interpolaciones suaves sucesivas entre dos tipos diferentes de una determinada letra. En *c* la letra de la izquierda es del tipo Bembo y la letra de la derecha es del tipo Negra Helvética; las formas intermedias han sido generadas por el programa Ikarus. Donald E. Knuth, de la Universidad de Stanford, ha desarrollado otro sistema de diseño conocido con el nombre de Metafont. Un programa escrito en lenguaje Metafont altera ciertas características de las letras controlando las propiedades de una pluma virtual; de acuerdo con ellas el ordenador representa la forma final de las letras en cuestión. Por ejemplo, se puede controlar la forma de la punta de la citada pluma virtual; dicha punta tiene la forma de una elipse en la que se debe especificar la relación entre los ejes horizontal y vertical, relación llamada aspecto (*d*). En esta ilustración se han representado también algunos ejemplos de los efectos de otras variables adicionales en el diseño de letras (*e, f, g*).

Juegos matemáticos

Tareas que es forzoso concluir, por mucho que se quiera evitarlo

Martin Gardner

Imaginemos una cesta que contiene 100 huevos, y que se dispone de suficientes envases de cartón. La tarea será colocar todos los huevos en cartones. Un paso, o movimiento, consiste bien en tomar un huevo de la cesta y colocarlo en un envase, bien en retirar un huevo de uno de los cartones y devolverlo al cesto. El procedimiento será éste: tras colocar uno tras otro dos huevos en cartones, es obligatorio devolver un huevo a la cesta. Salta a la vista que el procedimiento no es muy eficiente, pero no cabe duda de que todos los huevos del cesto acabarán embalados en cartones.

Supongamos ahora que la cesta pueda contener un número finito cualquiera de huevos. De ser lícito que el número inicial de huevos sea arbitrario, la tarea es no-acotada. Empero, tan pronto se concrete el número inicial, podrá determinarse una cota superior finita del número de pasos necesarios para concluir la tarea.

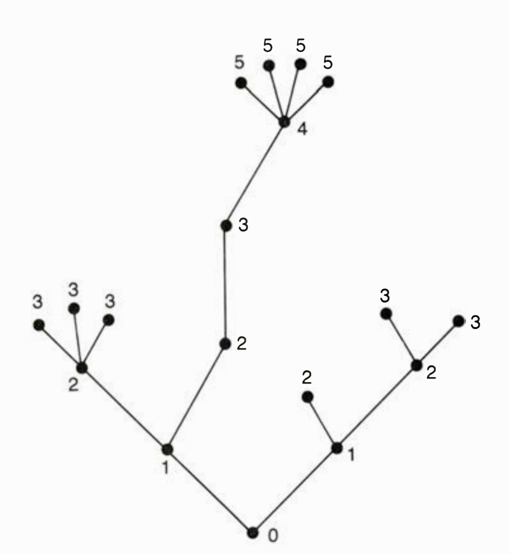
La situación cambia radicalmente si las reglas consienten transferir a la cesta cualquier número de huevos en cualquier momento que se quiera. Aunque la cesta contenga tan sólo dos huevos, ya no existe cota superior para el número de pasos requeridos para concluir la tarea. Según las reglas que se adopten, la tarea de colocar en los cartones un número finito de huevos puede tener forzosamente fin, puede ser imposible de terminar, o puede ser, a elección nuestra, de duración finita o infinita.

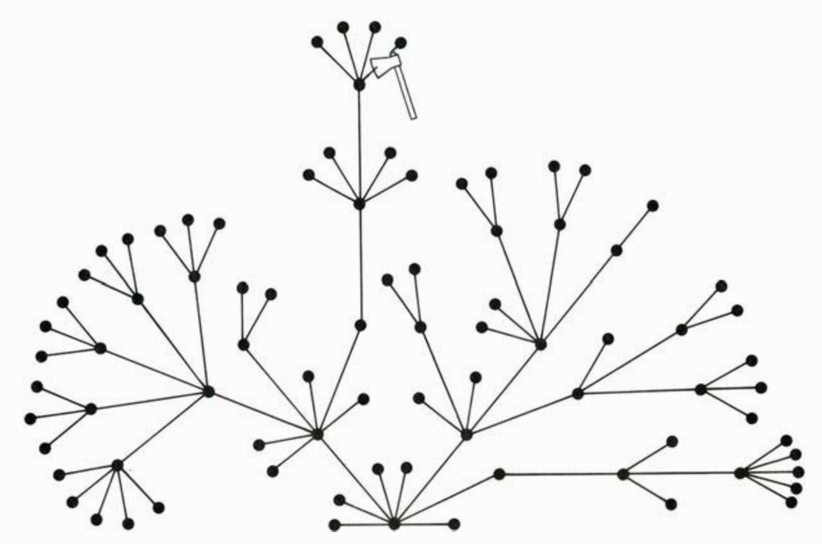
Nos ocuparemos ahora de varias amenas tareas matemáticas que comparten todas el siguiente rasgo: parece intuitivamente obvio que podremos demorar eternamente el fin de la tarea, mientras en realidad no hay forma de evitar concluirla en un número finito de pasos.

El primer ejemplo está tomado de un artículo del filósofo, escritor y lógico Raymond M. Smullyan. Imaginemos a nuestra disposición una reserva infinita de bolas de lotería, rotulada cada una con un número entero positivo, y que para cada entero existe también una colección infinita de bolas que lo portan. Tenemos además una caja que contiene una colección finita de bolas numeradas. El objetivo es vaciar la caja. Cada paso consiste en retirar una bola y reemplazarla con un número cualquiera positivo finito de bolas de rango inferior (rango de una bola es el valor grabado en ella). Las bolas de rango 1 son excepcionales, pues no existiendo bolas de rango inferior a 1, no serán reemplazadas por ninguna.

Es fácil vaciar la caja en un número finito de pasos. Bastaría, por ejemplo, reemplazar cada bola de rango mayor que 1 por una bola de rango 1, hasta dejar solamente bolas de este valor, y después ir sacando de una en una las bolas de rango 1. Sin embargo, las reglas nos permiten también reemplazar cada bola de rango mayor que 1 por *cualquier* número finito de bolas de rango inferior. Por ejemplo, podríamos retirar una bola de rango 1000 y sustituirla por 1000 billones de bolas de rango 999, más 100 billones de rango 998, más 10 billones de rango 997, y así sucesivamente. De esta forma, el número de bolas de la caja puede rebasar cualquier número imaginable en cada paso. ¿No será posible prolongar eternamente el contenido de la caja? Por increíble que al principio nos pueda parecer, no hay forma de evitar la terminación de la tarea.

Observemos que el número de pasos a dar para vaciar la caja es ilimitado, o no-acotado, en un sentido mucho más fuerte que en el juego de los huevos. No sólo no hay cota sobre el número inicial de bolas a vaciar, sino que, además, cada vez que se retira una bola de rango mayor que 1 tampoco hay cota para el número de bolas utilizadas para reemplazarla. Por tomar prestada una frase de John Horton Conway, el procedimiento es "ilimitadamente ilimitado". En cada etapa del juego, en tanto la caja contenga al menos una bola de rango mayor que 1, será imposible predecir cuántos pasos harán falta para vaciar la caja de todas ellas, exceptuadas las marcadas con 1. (Si todas las bolas son de rango 1, la caja quedará vacía, como es obvio, en tantos pasos cuantas bolas haya.) Empero, por muy sagaces y astutos que seamos al ir reemplazando las bolas, la caja acabará vacía tras un número finito de pasos. Evidentemente, es necesario admitir que, sin ser





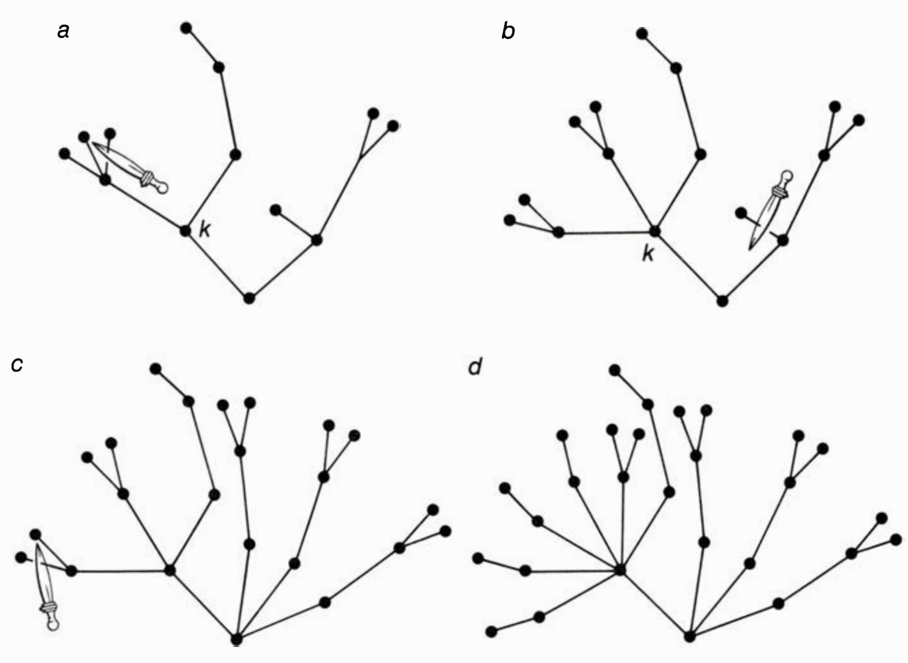
1. Una tarea de podadura

inmortales, viviremos lo bastante para completar la tarea.

Smullyan presenta este sorprendente resultado en un artículo titulado "Trees and Ball Games" ("Arboles y juegos de bolas"), aparecido en *Annals of the New York Academy of Sciences* (vol. 321, págs. 86-90; 1979). Smullyan ofrece allí varias demostraciones, incluido un sencillo razonamiento por inducción. Nada mejor que repetir aquí las palabras de Smullyan:

"Si todas las bolas de la caja son de rango 1 tenemos, como es obvio, un juego finito. Supongamos entonces que el más alto de los rangos de las bolas de la caja sea 2. Tendremos inicialmente un número finito de doses y un número finito de unos. No podemos estar eternamente sacando unos, y por tanto, más pronto o más tarde habremos de sacar alguno de los doses. Tendremos entonces un 2 menos en la caja (aunque posiblemente más unos que al comenzar). Como antes, no podremos estar eternamente sacando unos, así que llegará un momento en que sea forzoso sacar otro 2. Nos damos cuenta de que tras un número finito de pasos tendremos que eliminar nuestro último 2, con lo que estaremos de nuevo en el caso de tener solamente unos. Sabemos ya que es ésta una situación de pérdida, y esto demuestra que el proceso ha de terminar si el valor máximo de los rangos presentes es 2. Y ahora, ¿qué ocurrirá si el rango máximo es 3? No podremos estar sacando eternamente bolas de rango ≤ 2 (¡acabamos de probarlo!) y antes o después echaremos mano de un 3. Más adelante habremos de sacar un segundo 3, y con el tiempo, acabaremos sacando el último 3. El problema queda así reducido al caso precedente, en que el rango máximo era 2, y que ya hemos resuelto."

Smullyan demuestra también que el juego terminará con auxilio de una representación en grafo arbóreo. Por grafo arbóreo, o "árbol", entendemos un conjunto de segmentos rectilíneos, cada uno de los cuales une dos puntos, y ello de forma tal que exista un único punto, llamado "raíz del árbol" tal que cada punto del árbol quede conectado con la raíz a lo largo de un único sendero formado por dichos segmentos. El primer paso de estos juegos de bolas, llenar la caja, se expresa en el árbol representando cada bola por un punto de igual numeración que la bola, y uniéndolo por un segmento a la raíz del árbol. Cuando la bola es reemplazada por otras de rango inferior, se borra su número y las nuevas bolas son indicadas por un nivel más alto de puntos numerados, que se conectan con el punto ocupado por la bola eliminada. De esta forma el árbol crece constantemente hacia lo alto, con sus "puntos terminales" (puntos que como las hojas no están en la raíz, y están ligados a un solo segmento) representando siempre las bolas que ocupan la caja en esa etapa del juego.



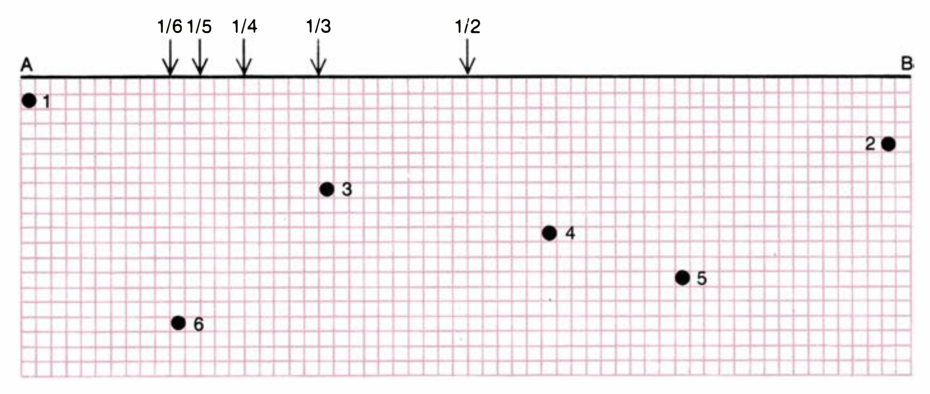
2. *Crecimiento de la hidra*

Smullyan demuestra que si este árbol llegase a hacerse infinito (esto es, tuviera una infinidad de puntos), habría de tener al menos una rama ascendente infinita. Pero esto es claramente imposible, porque los números de cada rama van decreciendo conforme se asciende por ella, y por consiguiente habrán necesariamente de terminar en 1. Puesto que el árbol es finito, el juego por él representado ha de terminar. Como sucedía en la versión con bolas, no hay manera de predecir cuántos pasos harán falta para terminar el árbol. En la etapa final, cuando el juego se convierte en un juego acotado, todos los puntos terminales están rotulados con 1. El número de los puntos de rango 1 puede, evidentemente, exceder del total de electrones del universo, o de cualquier otro número grande. Pero no es un juego de Sísifo: con certeza habrá de concluir en un número finito de jugadas.

El teorema básico de Smullyan, que él ha sido el primero en presentar con un modelo arbóreo, emana de teoremas concernientes a la ordenación de conjuntos que se remontan a trabajos de Cantor sobre números ordinales transfinitos, y está íntimamente emparentado con un profundo teorema sobre conjuntos infinitos de árboles finitos, demostrado inicialmente por Joseph B. Kruskal, y demostrado por segunda vez en forma más simple por C. St. J. A. Nash-Williams. Más recientemente, Nachum Dershowitz y Zohar Manna se han valido de razonamientos semejantes para demostrar que ciertos programas de ordenador, donde intervienen números "ilimitadamente ilimitados" de operaciones, han de llegar a detenerse.

Es posible construir un modelo de un caso particular de juegos de bolas de Smullyan numerando un árbol finito de la raíz hacia arriba, como vemos en la figura izquierda de la primera ilustración. Nos está permitido podar cualquier punto terminal finito junto con su segmento correspondiente, y añadir después las ramas nuevas que queramos al árbol, y hacer esto siempre que queramos, sin más condición que la de que los nuevos puntos sean de rango inferior al eliminado. Por ejemplo, la figura de la parte derecha de esa ilustración muestra uno de los posibles rebrotes, tras haber sido podado un punto de valor 4. A pesar de que tras cada poda puedan brotar en el árbol billones de nuevos retoños, al cabo de un número finito de podas el árbol habrá sido talado. A diferencia del juego general con bolas, no podemos eliminar puntos cualesquiera a capricho, sino tan sólo los terminales. Pero como cada punto eliminado se sustituye por otros de rango inferior, sigue siendo aplicable el teorema de Smullyan. Tras las podas, el árbol puede hacerse extraordinariamente frondoso, pero en cierto sentido va achaparrándose cada vez más, y termina por desaparecer.

Laurie Kirby y Jeff Paris han propuesto una forma más complicada de



3. Una forma de situar seis puntos sobre la recta **A-B**

talar un árbol en *The Bulletin of the London Mathematical Society* (vol. 14, 4.ª parte, n.° 49, págs. 285-293; julio de 1982). Llaman ellos “hidra” a su grafo arbóreo. Sus puntos terminales son las cabezas de la hidra, y Hércules quiere aniquilar al monstruo por decapitación total. Cada vez que se corta una cabeza queda eliminado también el segmento a ella ligado. Desdichadamente, después del primer tajo la hidra adquiere una o más nuevas cabezas, desarrollando una nueva rama a partir de un punto (llamémosle k) situado a un paso por debajo del segmento suprimido. Esta nueva rama es copia exacta de la parte de la hidra que, partiendo de k, crece hacia lo alto. La figura superior derecha de la segunda ilustración nos muestra la hidra después de que Hércules haya cercenado la cabeza indicada por la espada en la figura superior izquierda.

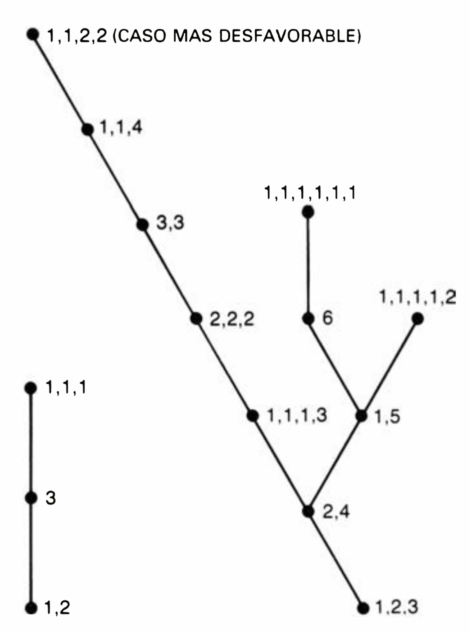
La situación de Hércules parece hacerse cada vez más desesperada, porque cuando corta la segunda cabeza, *dos* copias nacen justamente debajo del segmento cercenado (*figura inferior izquierda*). Y *tres* son las copias que brotan tras el tercer tajo (*figura inferior derecha*), y así sucesivamente. En general, tras el n-ésimo tajo rebrotan n nuevas copias. No hay manera de rotular los puntos de la hidra para que su crecimiento se corresponda con el juego de bolas de Smullyan; empero, Kirby y Paris consiguen demostrar, basándose en un notable resultado de teoría de números descubierto por el lógico y matemático inglés R. L. Goodstein, que, independientemente de la secuencia en que vaya Hércules segando cabezas, la hidra queda al fin reducida a un conjunto de cabezas (de las que puede haber millones, incluso siendo sencilla la forma inicial del monstruo) directamente unidas a la raíz. Son éstas eliminadas entonces una por una, hasta que, perdiéndolas todas, la hidra muere.

Una forma conveniente de atacar el juego de la hidra consiste en imaginar que el árbol representa un sistema de cajas encajadas (a modo de muñeca rusa). Cada caja contiene todas las cajas que podrían alcanzarse al ir ascendiendo por el árbol, y está rotulada con el número de niveles de encajonamiento que contiene. Así pues, en la primera figura de la hidra, la raíz es una caja de rango 4. Inmediatamente sobre ella tenemos una caja de rango 3 por la izquierda y otra de rango 2 por la derecha, y así sucesivamente. Todos los puntos terminales son cajas vacías, de rango 0 (cabezas de hidra). Cada vez que se elimina una caja de rango 0, la caja inmediatamente inferior queda duplicada (junto con todo su contenido), pero cada uno de los duplicados, lo mismo que la caja original, contiene ahora una caja vacía menos. Al cabo nos vemos obligados a comenzar a reducir los rangos de las cajas, como ocurría con las bolas del juego de Smullyan. Y un razonamiento inductivo análogo al explicado demostrará que finalmente todas las cajas han de quedar vacías, tras lo cual se eliminan una por una.

Debo esta explicación a Dershowitz, quien señaló que ni siquiera es necesario que la hidra limite su crecimiento a un número consecutivamente creciente de nuevas ramas; puede permitirse que tras cada tajo revivan, en número finito, tantas duplicaciones cuantas se desee. Tal vez ello le exija a Hércules mucho más tiempo para exterminar al monstruo, pero si continúa tajando sin parar, no hay manera de impedir la muerte de la hidra. Fijémonos en que, si bien la hidra va haciéndose más ancha, nunca va haciéndose más alta. Algunos de los más complicados programas crecientes analizados por Dershowitz y Manna pueden traducirse en grafos que no sólo se ensanchan, sino que ganan altura, y para tales árboles aún es más difícil demostrar que terminan.

Nuestro próximo ejemplo de tarea que aparentemente pudiera proseguir indefinidamente, pero que en realidad no es así, se conoce como problema de los 18 puntos. Se comienza con un segmento rectilíneo. Sobre él situamos, donde se quiera, un punto. A continuación se sitúa un segundo punto, pero con la condición de que cada uno de los dos puntos se halle en una mitad distinta del segmento. (Las mitades se consideran “intervalos cerrados”, pero los extremos no se consideran “interiores” a esa mitad.) Situemos un tercer punto de manera que cada uno de los tres se encuentre en distinto tercio del segmento. Ya salta a la vista que los dos primeros puntos no podrán tomarse en cualquier sitio. No pueden, por ejemplo, estar muy juntos, hacia el centro del segmento, o pegados ambos a un extremo. Deben encontrarse cuidadosamente situados, de modo que cuando se añada un tercer punto cada uno de los tres se halle en diferente tercio del segmento. Hay que proseguir de igual manera, colocando el n-ésimo punto de modo que los n primeros puntos ocupen siempre diferentes $1/n$-ésimas partes iguales del segmento. Si se eligen cuidadosamente las posiciones de los puntos, ¿cuántos puntos podremos situar sobre el segmento?

Parece, intuitivamente, que tal número debiera ser ilimitado. Evidentemente, un segmento puede dividirse en tantas partes iguales como se quiera, y cada una de ellas puede siempre contener un punto en su interior. La dificultad está en que para cumplir la condición los puntos han de estar suce-



4. Arboles correspondientes al solitario, con tres y seis naipes

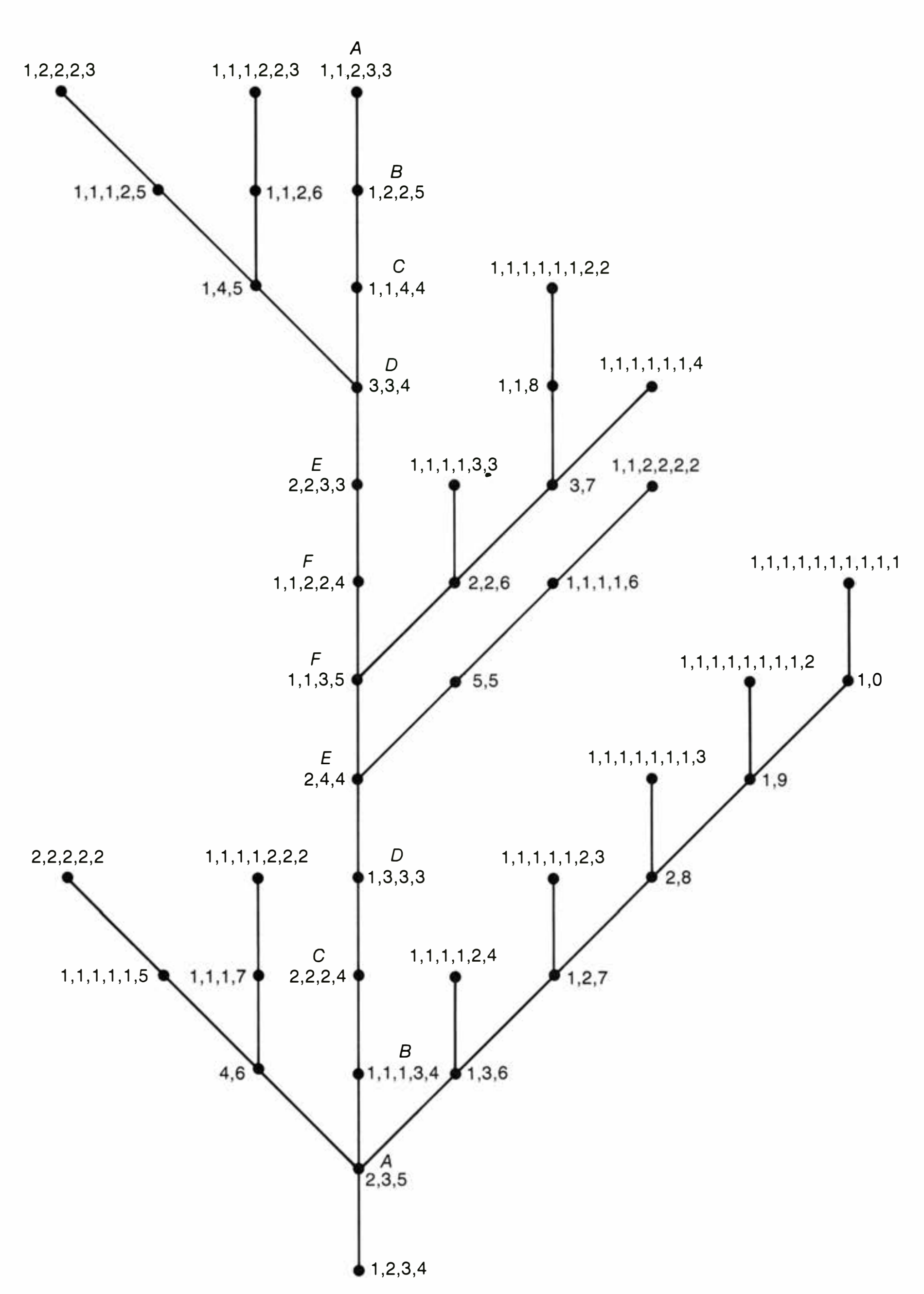
sivamente numerados. ¡Resulta imposible pasar de 17 puntos! Por muy sagaz que se sea al ir colocando los 17 puntos, el decimoctavo tiene que violar la regla, y el juego terminará. La verdad es que no es fácil colocar siquiera 10 puntos. La figura 3 nos enseña un procedimiento para colocar seis.

Este insólito problema apareció por primera vez en *One Hundred Problems in Elementary Mathematics* (problemas 6 y 7) del matemático polaco Hugo Steinhaus. Steinhaus da una solución con 14 puntos, y, en nota al pie, añade que M. Warmus ha demostrado que el límite es 17. La primera demostración, publicada por Elwyn R. Berlekamp y Ronald L. Graham, se encuentra en su artículo "Irregularities in the Distributions of Finite Sequences", *Journal of Number Theory* (vol. 2, n.º 2, págs. 152-161; mayo de 1970).

Warmus, un matemático de Varsovia, no publicó su demostración, más breve, hasta seis años más tarde, en la misma revista (vol. 8, n.º 3, págs. 260-263; agosto de 1976). Da una solución de 17 puntos, y añade que hay 768 disposiciones de la solución, o 1536 si se consideran distintas las que tienen sus puntos en orden inverso.

Nuestro último ejemplo de tarea que concluye súbitamente y de forma inesperada a la intuición se presta bien a traducirlo a un juego de cartas. Su origen es desconocido, pero Graham, que me lo refirió, dice que los matemáticos europeos lo llaman "el solitario búlgaro", por razones que no alcanzó a descubrir. Como es bien sabido, las sumas parciales de la serie 1 + 2 + 3 + ... son los llamados números triangulares, porque corresponden al número de elementos de formaciones triangulares tales como los bolos o las bolas de billar de troneras (10 y 15, respectivamente). La tarea concierne a cualquier número triangular de naipes. En una baraja de 52 cartas, el máximo número triangular que podemos extraer es 45, suma de los nueve primeros números naturales.

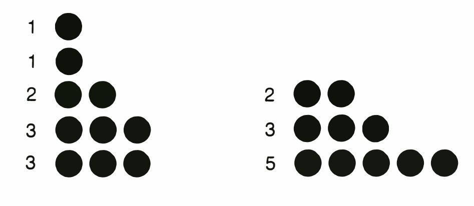
Formemos un mazo con las 45 cartas, y dividámoslo luego en cuantas pilas queramos con número arbitrario de cartas en cada pila. Podemos dejarlo formando una única pila de 45 naipes, o cortarlo en dos, o en tres, o en más montones, por donde se quiera; en el caso extremo, dando 44 cortes tendríamos 45 pilas de una carta cada una. Ahora, repitamos una y otra vez el siguiente procedimiento: se toma una carta de cada montón, y con todas ellas se forma un nuevo montón sobre la mesa. No es preciso que los montones estén alineados; podemos situarlos a placer. Repitamos el procedimiento, formando otra pila, y así una y otra vez.



5. Arbol búlgaro para 10 cartas

Conforme la estructura de las pilas va cambiando irregularmente, parece que será improbable que llegue a alcanzarse una situación en la que haya exactamente una pila con un naipe, una pila con dos, una con tres, y así sucesivamente, hasta una con nueve. Si llegase a crearse semejante estado, sin quedar antes atrapados en ciclos que hagan retornar el sistema a estados anteriores, es preciso que el juego se dé por terminado, porque ahora su estado no puede variar. La repetición del procedimiento deja a los naipes exactamente con la misma distribución consecutiva que antes tenían. Y para nuestra sorpresa, resulta que, independientemente del estado inicial del juego, es inexorable que se alcance el estado de montones consecutivos en número finito de pasos.

El solitario búlgaro nos da un procedimiento para construir modelos de ciertos problemas de la teoría de particiones, problemas que distan de ser triviales. Las particiones de un número natural n son todos los modos en que el número n puede expresarse como suma de enteros positivos, sin tener en cuenta el orden de los sumandos. Por ejemplo, el número triangular 3 tiene tres particiones, a saber, 1 + 2, 1 + 1 + 1, y 3. Al dividir el mazo de naipes en un número arbitrario de pilas, lo que estamos haciendo es una partición del número de cartas del mazo. El solitario búlgaro es un procedimiento para



6. Los pares conjugados 1,1,2,3,3 y 2,3,5

transformar unas particiones en otras, a base de restar uno de cada número de la partición y añadir luego otro sumando igual al número de unos sustraídos. No es obvio que este procedimiento siempre vaya a dar lugar a una cadena de particiones libre de repeticiones, y que termine produciendo finalmente la partición consecutiva. Me han dicho que el primero en demostrarlo fue Jørgen Brandt, un matemático danés, en 1981; pero no conozco su demostración ni sé tampoco si ha sido publicada.

El solitario búlgaro de un número triangular cualquiera de naipes puede transformarse en un grafo arbóreo cuya raíz esté rotulada con la partición consecutiva, y todas las demás particiones estén representadas por puntos del árbol. La figura izquierda de la cuarta ilustración muestra el sencillo árbol correspondiente al juego de tres naipes; la figura de la derecha es el árbol, no tan trivial, formado por las 11 particiones de seis naipes. El teorema que afirma que todo juego del solitario terminará en la partición consecutiva es equivalente al teorema de que todas las particiones de un número triangular admiten como modelo un grafo arbóreo conexo donde cada partición se encuentre un paso por encima de su siguiente en el juego, y donde la partición consecutiva esté plantada en la raíz del árbol.

Observemos que el punto más elevado del árbol correspondiente a seis cartas se encuentra a seis pasos de la raíz. Esta partición, 1, 1, 2, 2 supone el caso inicial más desfavorable. Es fácil comprender que el juego debe terminar en no más de seis pasos, cualquiera que sea la posición de partida. Se ha conjeturado que cualquier juego ha de terminar en no más de $k\ (k-1)$ pasos, donde k es cualquier número entero positivo y el número triangular es $\frac{1}{2}k(k+1)$. El pasado año, Donald E. Knuth, profesor de ciencias de cómputo en la Universidad de Stanford, pidió a sus alumnos que verificasen la conjetura mediante ordenador, logrando confirmarla para valores de k menores o iguales que 10, con lo que la conjetura es muy verosímil; pero hasta ahora la demostración no se ha dejado atrapar.

La figura 5 muestra el árbol correspondiente al solitario búlgaro con 10 naipes (k=4). Tenemos en lo alto los tres casos más desfavorables, que distan cada uno 12 pasos de la raíz. Observemos también que el árbol tiene 14 puntos terminales. Los llamaremos "particiones del Edén", porque, a menos que se comience con alguna de ellas, jamás se presentan durante el desarrollo del juego. Son todas aquellas particiones cuyo número de partes excede al más alto número de partes en dos o más.

La figura izquierda de la sexta ilustración muestra la forma en que ordinariamente se utilizan puntos para representar particiones gráficamente. La que vemos corresponde a 1, 1, 2, 3, 3, situada en la cúspide del árbol. Haciéndola girar 90 grados y trasponiéndola por simetría, esta formación de puntos se convierte en la dibujada a su derecha, cuyas filas dan ahora la partición 2, 3, 5. Cada una de estas particiones se llama "conjugada" de la otra. La relación es obviamente simétrica. Las particiones que, sometidas a conjugación, quedan invariables se llaman autoconjugadas. En el árbol de 10 naipes hay exactamente dos de tales particiones: la raíz, y la 1, 1, 1, 2, 5. Al emparejar por conjugación las restantes particiones se presenta a lo largo del tronco una curiosa pauta. Las particiones se emparejan como muestran las letras. Esta simetría se presenta a lo largo del tronco principal de todos los "árboles de Bulgaria" hasta ahora examinados.

Si la simetría observada fuese válida para todos los árboles de este tipo tendríamos una forma sencilla para determinar el peor de los casos que pueden presentarse en la cima: es el conjugado de la partición (siempre hay exactamente una) directamente situada sobre la raíz. Un procedimiento aún más rápido para hallar la cúspide del tronco es añadir a la raíz el prefijo 1, y disminuir en una unidad su último número.

La "operación búlgara" puede traducirse a diagrama, suprimiendo la columna de la izquierda en la formación de lunares, girando 90 grados la columna eliminada, y añadiéndola como nueva fila. Tan sólo los diagramas de la forma 1, 2, 3, 4, ... quedan invariables al ser sometidos a este proceso. Si pudiésemos demostrar que ninguna secuencia iterada de operaciones podrá devolver el diagrama a su estado primitivo, se habría demostrado que todos los juegos búlgaros se traducen en árboles, y deben por consiguiente terminar cuando se alcance su raíz.

Cuando el juego se efectúa con 55 naipes (k = 10) hay 451.276 formas de particionarlos, y difícil resultaría dibujar el árbol. Incluso el árbol de 15 cartas, con sus 176 puntos, requiere ayuda de ordenador. ¿Pero, cómo se han calculado estos números? Se trata de una historia larga y fascinante. Pongamos que las particiones tuvieran en cuenta el orden de los sumandos, con lo que 3, por ejemplo, tendría cuatro particiones ordenadas (por lo general denominadas "composiciones"): 1 + 2, 2 + 1, 1 + 1 + 1, 3. Resulta que la fórmula del número total de composiciones es, sencillamente, 2^{n-1}. Pero cuando las particiones están desordenadas, como sucede en el solitario, la situación se hace increíblemente más confusa. Y aunque hay muchos métodos iterativos o recursivos para contar particiones desordenadas, que se valen en cada paso del número de particiones conocidas para números más pequeños, no pudo lograrse una fórmula asintótica exacta hasta tiempos recientes. El gran progreso se debió al matemático inglés G. H. Hardy, en colaboración con su amigo indio Srinavasa Ramanujan. Su fórmula asintótica, no totalmente exacta, fue perfeccionada por Hans A. Rademacher en 1937. La fórmula de Hardy–Ramanujan–Rademacher es una serie infinita de aterradora apariencia, que se vale, entre otros ingredientes, ¡del número pi, de raíces cuadradas, raíces de complejos y derivadas de las funciones hiperbólicas! George E. Andrews, en su clásico texto sobre teoría de particiones, dice de la fórmula que es "una identidad increíble" y que es "uno de los logros cimeros" de la historia de su materia.

La sucesión de particiones para n=1, n=2, n=3, n=4, n=5 y n=6 es 1,2,3,5,7,11, y parece que se debería esperar que el siguiente número de particiones fuese el número primo siguiente, 13. Lástima, porque es 15. Tal vez todas las particiones sean impares. Pues no; la partición siguiente es 22. Uno de los problemas más profundos de la teoría de particiones, pendiente de solución, es el de averiguar si cuando n crece, el número de particiones pares y el de particiones impares tienden a igualarse.

Tal vez le parezca al lector que la teoría de particiones es poco más que un diletantismo de matemáticos. Permítame en tal caso terminar diciendo que una cierta forma de representar mediante diagramas conjuntos de particiones, con ayuda de tablas de números llamadas tablas de Young, ha llegado a ser enormemente útiles en la física de partículas. Pero ese es otro juego de bolas.

Taller y laboratorio

De cómo fórmulas sencillas informan de si un nudo aguantará o se soltará

Jearl Walker

Un nudo es un lazo que se estrecha y cierra con una o más cuerdas o cabos para unirlos entre sí o para sujetar alguna cosa. ¿Hay nudos más seguros que otros? ¿Tiene algún límite la carga que puede soportar un nudo antes de que se deshaga porque su extremo libre se escurra por entre el lazo? ¿Hay manera de modificar un nudo para que resista una carga mayor? Para discurrir en torno a estas preguntas me servirán de pauta las investigaciones de Benjamin F. Bayman, de la Universidad estatal de Minnesota, autor de una impresionante teoría acerca de las resistencias de los nudos.

La seguridad de un nudo depende primordialmente de dos cosas: del modo eń que se arrolla alrededor de un objeto (en lo que yo voy a llamar volteos) y del modo en que se cruza sobre sí mismo (cruces). Ambos son responsables del rozamiento que se desarrolla para mantener el nudo cuando la cuerda se somete a tracción. Sin ese rozamiento la cuerda resbalaría en toda la lazada y el nudo se soltaría. Me limitaré a cuerdas de diámetro mediano atadas en torno a barras o postes de mayor diámetro. ¿Qué grado de rozamiento en los volteos y cruces puede asegurar que un nudo no va a fallar, aun cuando la cuerda se someta a una carga arbitrariamente grande?

Pasaré por alto algunos aspectos prácticos de los nudos: ¿por qué algunos son más adecuados para cargas que tiren perpendicularmente a la barra? ¿por qué usar otros que tiren paralelamente a la misma? Los hay que ofrecen ciertas ventajas por el hecho de aguantar mucho tiempo, aunque estén deshaciéndose. Ciertos tipos trabajan mejor si la intensidad de la carga varía periódicamente. En nuestro caso, todos esos factores son secundarios. Tampoco voy a tener en cuenta la resistencia adicional que puede ganar un nudo en virtud del modo en que froten entre ellas porciones de cuerda adyacentes (salvo en los cruces). Los orígenes principales del frotamiento en los nudos son, para nosotros, los volteos y los cruces.

Para comprender la función que la tracción desempeña en una cuerda sometida a una carga imaginemos un peso pendido de una cabo. Que la cuerda está sometida a tracción significa que cualquier porción de la misma estará en equilibrio bajo la acción de dos fuerzas que tiran en sentidos contrarios desde las dos porciones contiguas. En toda la longitud de la cuerda. la tracción es constante. Por tanto, para que el conjunto permanezca inmóvil, la fuerza aplicada en el extremo superior de la cuerda debe igualar al peso sujeto al extremo inferior.

Imaginemos ahora que sujetamos el peso pasando la cuerda sobre una barra horizontal fija. Entonces el trozo de cuerda comprendido entre la barra y el peso estará sometido a una tracción constante; el trozo en contacto con la barra sufre, sin embargo, una tracción variable, que decrece desde la sección contigua al trozo del que cuelga el peso hasta la sección en la que la cuerda se separa de la barra. La tracción en el trozo comprendido entre esta última sección y nuestra mano será constante, pero inferior a la del trozo situado al otro lado de la barra, a causa del rozamiento entre ésta y la cuerda. O sea, como la cuerda frota con la barra ya no hay que aplicar la misma fuerza para sostener el peso.

El valor de la fuerza de rozamiento que actúa sobre la cuerda depende parcialmente de la aspereza de las dos superficies en contacto y, también, de la longitud de cuerda que roza con la barra. Es posible aumentar el rozamiento dando varias vueltas a la cuerda en torno a la barra, en una maniobra que montañeros y marinos conocen como encabillar. Cada vuelta más reducirá la fuerza que hemos de aplicar en nuestro extremo de la cuerda.

En una lazada puede asimismo generarse rozamiento cruzando la cuerda sobre sí misma. Cuando la cuerda se somete a tracción, la parte superior del cruce comprime a la parte inferior contra la barra. Esta compresión se opone a la tendencia a deslizar de la cuerda, de tal modo que cuanto más fuerte sea la compresión tanto más intenso será el rozamiento. Así, un nudo no se deshará siempre que el frotamiento sea tan intenso para contrarrestar la fuerza que trata de que la parte de la cuerda que está debajo se deslice por entre el cruce.

Los nudos tienen por misión que el rozamiento que se genera dentro de la lazada resista la tracción que genere la carga. Supongamos que ésta aumente poco a poco. ¿Acaba alcanzando la carga un valor para el cual el extremo libre de la cuerda comience a escurrirse por el nudo, o bien el aumento de tracción se limitará a reforzar el rozamiento de suerte que la cuerda no resbale lo más mínimo? Si acontece que el rozamiento aumenta automáticamente, el nudo resulta ser autoapretante y no podrá ceder bajo carga alguna; antes se romperá la cuerda. Que un nudo se suelte o resista depende de su topología y del rozamiento producido por sus volteos y cruces.

Consideremos un volteo que sea múltiplo entero de una vuelta completa. A un lado de la cuerda la tracción es alta: sea ésta T_2. En el otro extremo la tracción será menor: sea ésta T_1. Esa cuerda no deslizará si T_2 no es mayor que T_1 multiplicada por cierto coeficiente, coeficiente que es función exponencial de dos cantidades: el ángulo (en radianes) de arrollamiento de la cuerda en torno a la barra y el coeficiente de rozamiento entre la cuerda y la barra. Así, si la cuerda da una vuelta en torno a la cuerda, el coeficiente será la función exponencial de 2π radianes multiplicado por el coeficiente de rozamiento. Para simplificar la función la escribiremos e. Entonces, si T_2 no excede de $e \times T_1$, la cuerda no resbalará. Pero si T_2 supera ese valor, el rozamiento en-

tre la barra y la cuerda será desbordado y ésta resbalará.

Si la cuerda da dos vueltas alrededor de la barra, el coeficiente será la función exponencial de 4π radianes por el coeficiente de rozamiento. Esta función podemos escribirla de manera sencilla como e^2. La cuerda no resbalará cuando T_2 no exceda el producto $e^2 \times T_1$. Esta relación muestra la ventaja de toda vuelta adicional que se dé a la cuerda. Así T_2 puede ser mucho mayor que T_1 pues el rozamiento que aporta la vuelta adicional contribuye a sostener la cuerda.

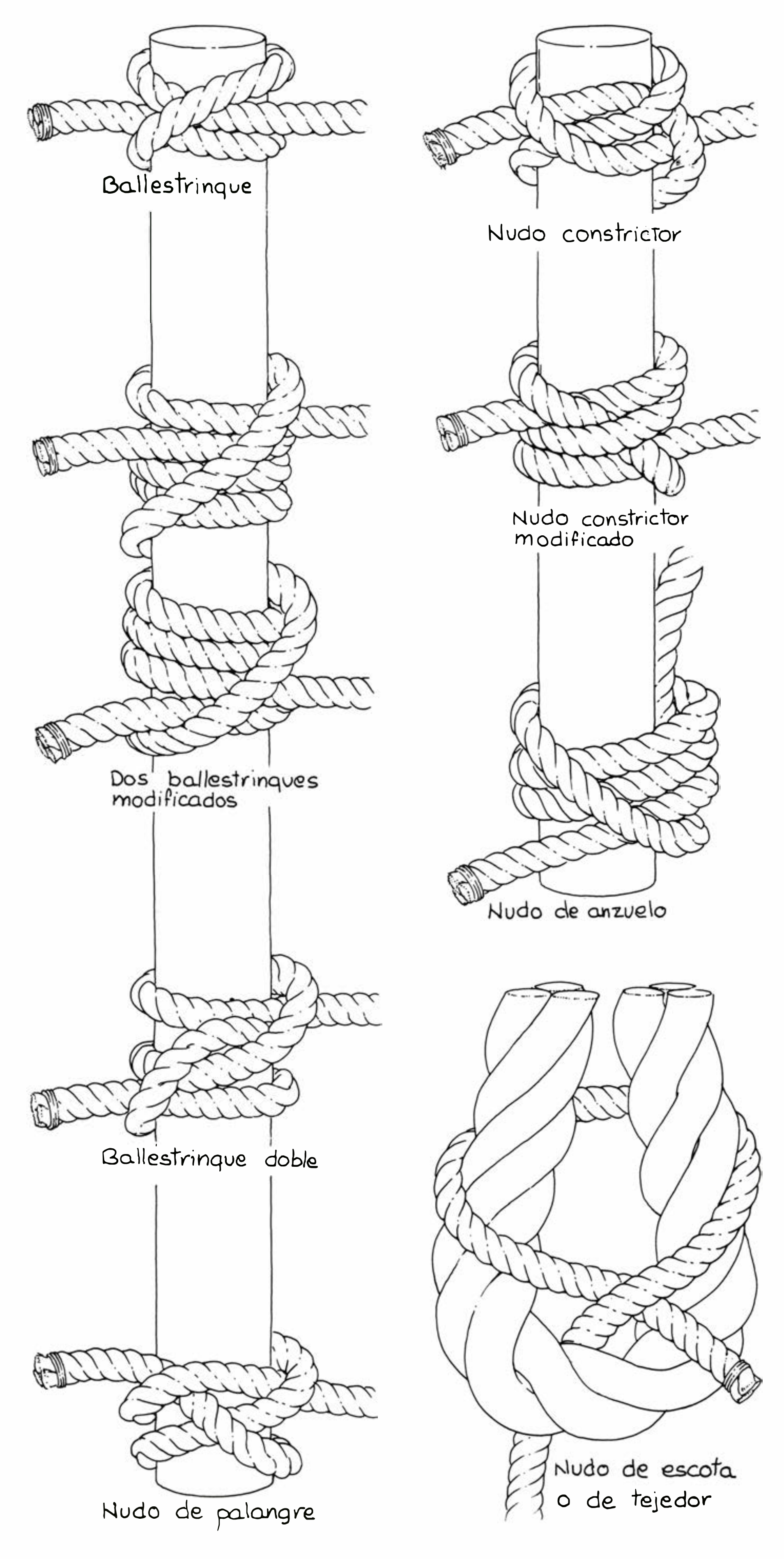
Para todo número entero de vueltas de la cuerda alrededor de la barra, la relación entre las tracciones en ambos extremos de la cuerda presenta una forma similar. En general, si es n el número de vueltas, la cuerda no resbalará mientras T_2 no sea mayor que $e^n \times T_1$.

El otro procedimiento para incrementar el rozamiento en un nudo se basa en los cruces. En éstos, la parte superior comprime, contra la barra, la parte inferior y, así, ésta se resistirá a deslizarse por dentro del cruce y aflojar el nudo. Consideremos las tracciones en los dos lados de la parte inferior de un cruce genérico. En el lado donde la tracción es mayor (T_2), la fuerza tendente a que deslice por entre el cruce la parte inferior de la cuerda valdrá T_2.

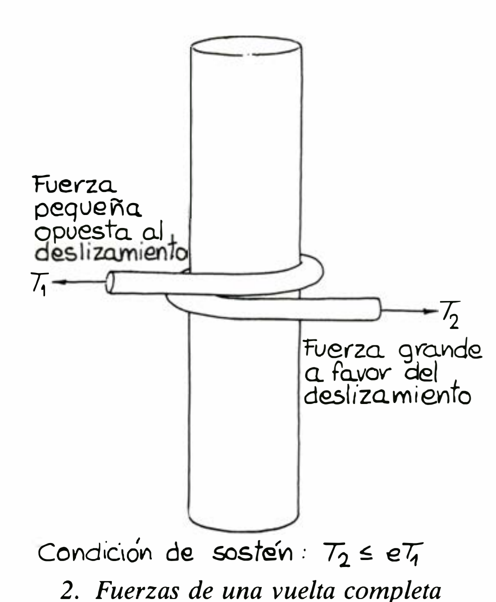
Al otro lado del cruce se tiene la tracción T_1, de menor valor, que se opone a dicha tendencia. Pero un factor adicional que se opone también a esa tendencia es el rozamiento debido a la compresión ejercida por el cruce. A lo sumo ese rozamiento puede valer la tracción en la parte superior del cruce multiplicada por cierto coeficiente (u). Así, la parte inferior de la cuerda no se deslizará dentro del cruce mientras T_2 no exceda de $T_1 + uT$. Si T_2 es demasiado grande, el rozamiento que aporta la parte superior no bastará para mantener en su sitio a la parte inferior.

El valor de u no es fácilmente calculable para un caso general, pues en él intervienen el coeficiente de rozamiento entre la barra y la cuerda de abajo y el coeficiente de rozamiento entre los trozos de cuerda sobrepuestos. Depende, además, de los diámetros de barra y cuerda, elemento geométrico éste que afecta a la aplicación de la fuerza que ejerce el trozo de arriba sobre el de abajo. Bayman ha medido u y e para lazos hechos sobre barras de acero pulido con cordeles de nylon trenzado; e resultó ser 4, aproximadamente, y u del orden de 0,2.

En la ligadura que los marineros lla-



1. Nudos diversos



2. Fuerzas de una vuelta completa

man ballestrinque intervienen volteos y cruces. Se trata de un nudo común que permite asegurar un cabo a una pieza fija o a una soga más gruesa. Partiendo del extremo libre, la cuerda da una vuelta a la barra para regresar hacia atrás. Seguidamente se arrolla otra vez en torno a la barra antes de salir desde debajo de un segundo cruce.

En la figura 4 se rotulan los tramos de cuerda, empezando con 0 para el extremo libre. El tramo 1 comienza en el primer cruce y continúa hasta que la cuerda pasa por debajo del segundo cruce. El tramo 2 sale del segundo cruce en dirección a la carga. (En la realidad el nudo está más tenso que en la figura.)

Veamos ahora. ¿Es éste un nudo eficaz? ¿Qué ocurre con la carga en 2? ¿Puede crecer hasta cualquier valor o acabará forzando al extremo libre a deslizarse a través del primer cruce y a deshacer todo el nudo? Consideremos la porción de abajo del primer cruce. A uno de sus lados, el cabo suelto no sufre tracción. En el otro, la cuerda está sometida a una tracción T_1 que actúa tratando de que el extremo libre resbale por dentro del cruce. A esto, la única fuerza que se opone es el frotamiento que genera la parte superior del cruce en el punto donde oprime a la parte inferior contra la barra. Para que el nudo resista, T_1 no puede superar el valor máximo posible de ese frotamiento.

Consideremos ahora las dos vueltas del tramo 1. Estas no resbalarán si la tracción en la parte contigua al segundo cruce no excede de $e^2 \times T_1$. Por último, tenemos el segundo cruce. Por debajo de éste la cuerda no resbalará si la tracción T_2 producida por la carga no sobrepasa las fuerzas que a ella se oponen en el cruce.

Para que el nudo aguante debe satisfacerse el sistema de inecuaciones constituido por esas relaciones entre fuerzas en distintas partes del nudo. Este sistema puede resolverse por sustitución, al igual que un sistema de ecuaciones corriente. El resultado es tajante: si u es mayor que $1/e$ se satisfacen todas las condiciones que imposibilitan el resbalamiento. Si se cumple esta sencilla condición, la carga sobre el nudo puede ser arbitrariamente grande, pese a lo cual el nudo seguirá resistiendo. Toda carga adicional aumentará los esfuerzos tendentes a aflojar la lazada, pero aumentarán asimismo las fuerzas de tracción generadoras del rozamiento antagonista.

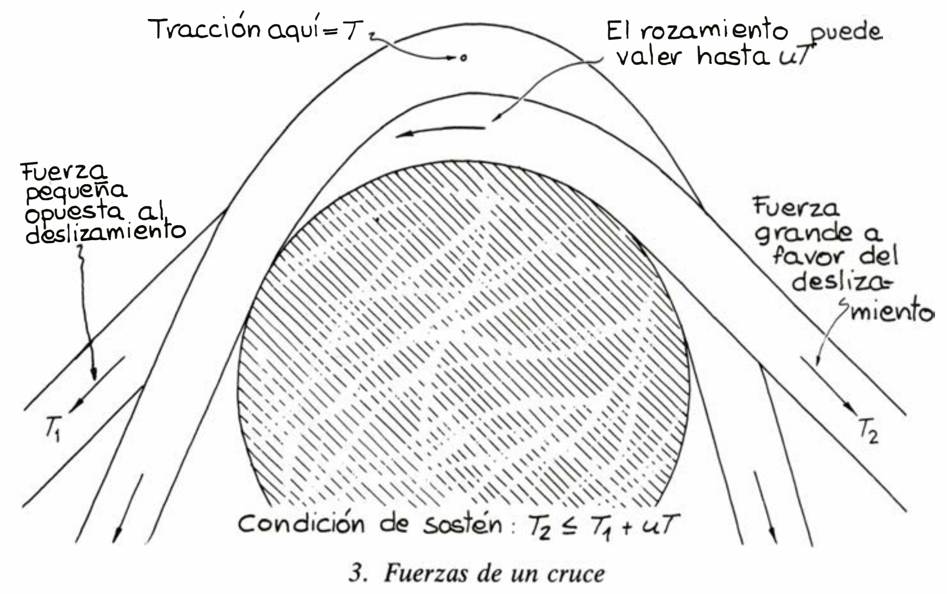
En un ballestrinque el elemento crítico lo encontramos en el primer cruce. En éste, la fuerza que tira del extremo libre tratando de hacerlo pasar por debajo del mismo vale T_1. La fuerza que a ello se opone vale, como mucho, u por la tracción en la cuerda que pasa por encima del cruce. Esta tensión se relaciona sin dificultad con T_1 porque entre los dos puntos de la cuerda hay una vuelta sólo. Como la tracción en la parte de arriba de la cuerda es eT_1 como máximo, el rozamiento en el cruce será de ueT_1, como máximo.

¿Mejorarían un ballestrinque algunas vueltas más, en el sentido de que resista cualquier carga aunque u sea más pequeño? En la figura 6 se representan dos ballestrinques modificados. En el primero de ellos la cuerda se arrolla dos veces más en torno a la barra antes de que cruce por encima del límite entre los tramos 0 y 1. En el segundo, las dos vueltas adicionales se dan inmediatamente antes de que la cuerda pase por debajo de su último cruce en dirección a la carga. Esos nudos son, desde luego, más complicados que el ballestrinque normal; pero, ¿son más fuertes por depender menos del valor de u? ¿Es uno de ellos mejor que el otro?

El elemento clave en estos ballestrinques modificados se encuentra en la pugna que se establece junto al extremo libre. Consideremos al primero de ellos. El tramo 1 trata de que el extremo libre pase por debajo del cruce aplicando una tracción T_1. Pero la cuerda de la parte superior del cruce se opone, apretando la cuerda.

¿Deslizará la cuerda? No, si u es suficientemente grande. ¿Debe ser éste del mismo valor que en un ballestrinque normal? No, puede ser más pequeño gracias a las vueltas adicionales. La tracción en la parte superior del primer cruce es ahora mucho mayor que la fuerza (T_1) tendente a hacer que la cuerda deslice por debajo del primer cruce. En un ballestrinque normal esa fuerza debida al cruce puede valer ueT como máximo. El exponente 1 (implícito) de e representa que la cuerda da una sola vuelta desde la parte de arriba del cruce y el punto en que T_1 tira de la parte inferior del mismo. En este ballestrinque modificado, la fuerza debida al cruce puede valer hasta ue^3T_1. Aquí el exponente 3 (explícito) de e señala los tres volteos que abarca el cruce. El nudo modificado tiene, pues, más posibilidades de mantenerse firme.

¿Qué decir de la segunda modificación? Parece tan complicada como la primera, pero aquí, también, la clave estriba en asegurar el extremo suelto.



3. Fuerzas de un cruce

Esta vez la pugna de esfuerzos es exactamente igual que en un ballestrinque normal. Las vueltas suplementarias no nos sirven para afianzar el extremo libre y, por ello, no suponen mejora alguna respecto al ballestrinque normal. Lo que si hacen es aumentar el frotamiento sobre la cuerda, reduciendo así el valor de T_1 cuando la cuerda se somete a carga; mas, para que aguante el nudo, u debe seguir siendo mayor que $1/e$ (como en el ballestrinque normal).

Del primer ballestrinque modificado podemos sacar alguna otra conclusión. Para que resista, u debe ser mayor que $1/(e + e^2 + e^3)$. Empero, si u es mayor que $1/(e + e^2)$, la primera vuelta no sirve para nada y puede eliminarse sin poner en peligro la integridad del nudo. El tramo 0 podría muy bien eliminarse pasándolo por debajo del cruce, con lo que se integraría en el tramo 1, en parte del cual se convertirá. Más aún, si u es mayor que $1/e$, el tramo 2 no sirve para nada y puede también suprimirse. O sea, si u es mayor que $1/e$ no hay razón para añadir vueltas a un ballestrinque: el nudo no se deshará.

Otra modificación posible es el ballestrinque doble ilustrado en la figura 7. Este nudo difiere del ballestrinque normal en que, después de que la cuerda pasa por debajo del segundo cruce, no se encamina directamente a la carga, sino que da otra vuelta y hace otro cruce.

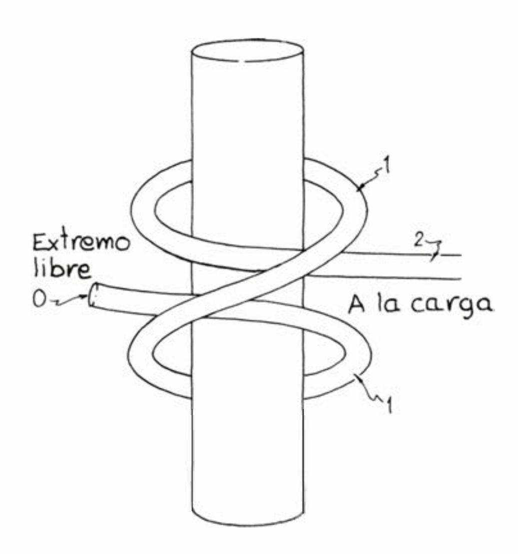
¿Reducen estas aditamentos el valor requerido a u? No; al menos no con las hipótesis que hemos hecho acerca de los orígenes principales del rozamiento en los nudos. Para que el ballestrinque doble resista, u tiene que seguir siendo mayor que $1/e$. El anudamiento suplementario sirve primordialmente de elemento de seguridad. En efecto, si el tramo 0 deslizara en su cruce, aún quedaría un ballestrinque que tendría que desatarse antes de que la cuerda se desprendiera por completo de la barra.

Un nudo similar al ballestrinque es el nudo de palangre que se muestra en la figura 8. Difiere del primero en el modo en que se anudan sus primeros tramos. El tramo 0 pasa bajo un cruce debido a un tramo final del nudo, es decir, a un tramo contiguo al que soporta toda la carga. El tramo 1 da una vuelta completa a la barra y luego pasa bajo sí mismo para iniciar el tramo 2, el cual, a su vez, se arrolla una vez en torno a la barra antes de pasar bajo el último cruce e iniciar el tramo 3.

El estudio del sistema de desigualdades revela que este nudo aguanta si u y e cumplen la condición de que $ue\,(u+e)$ sea mayor que 1. Como e es mayor que 1, dicha condición se cumplirá para un valor de u pequeño, inferior al de $1/e$ que pide el ballestrinque. O sea, hay una gama de valores de u para el cual el nudo de palangre resiste, pero no el ballestrinque. Si u es mayor que $1/e$, ambos nudos sirven lo mismo.

¿Cómo logra un nudo de palangre aguantar con un u insuficiente para sostener el ballestrinque? La respuesta está en el modo en que se remete el cabo libre. Este, en el ballestrinque, se encuentra bajo el tramo de cuerda 1, después que el mismo ha dado una vuelta. Pero en el nudo de palangre está remetido bajo un tramo de la cuerda más próximo a la parte bajo plena carga. Por tanto, la tracción en la parte superior del cruce, la que inmoviliza el extremo libre, es mayor que en el ballestrinque. En general, todo nudo será más firme (en el sentido de que u puede ser menor) cuando el extremo libre se remete de esa manera.

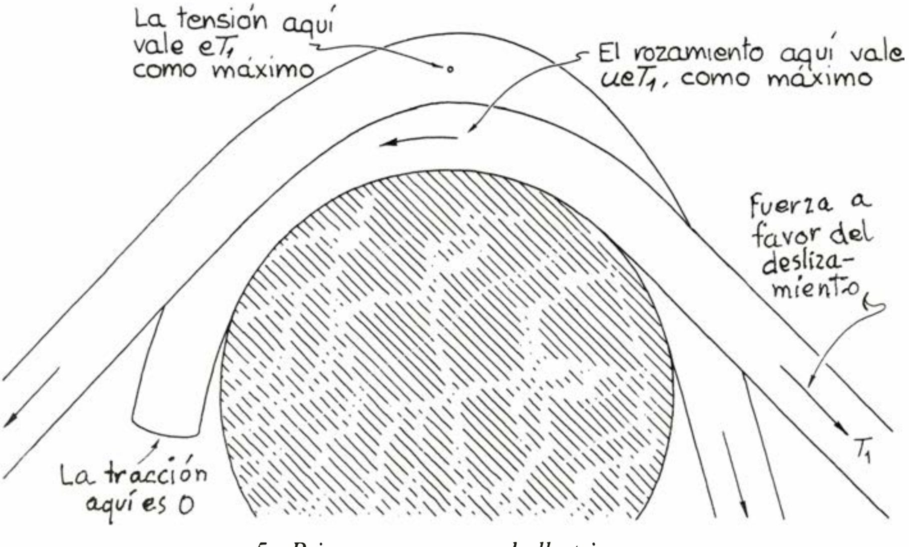
Bayman presenta otro ejemplo de cómo remeter el extremo libre provechosamente. En el nudo constrictor el extremo libre discurre bajo dos cruces, cada uno de los cuales sufre una gran tracción. Resolviendo el sistema de inecuaciones se consigue el valor límite de u para que el nudo no se deshaga: u debe ser mayor que $1/(e(2 + e))$. Como esta condición es menos rigurosa que la del ballestrinque, resulta que el nudo constrictor es mejor desde el punto de vista del valor de u. Si éste es mayor que $1/(2e)$, el rozamiento en el cruce que separa los tramos 1 y 2 será suficiente para que la tracción en el tramo 1 pueda ser nula sin que resbale la cuerda. Si u es mayor que $1/e$, ya no hará falta el tramo 2, e incluso puede ser nula la tracción en ella sin que peligre el nudo. Por ello, un nudo constrictor completo sirve perfectamente cuando u es pequeño comparado con $1/e$, pero no es necesario que sea completo para valores mayores de u.



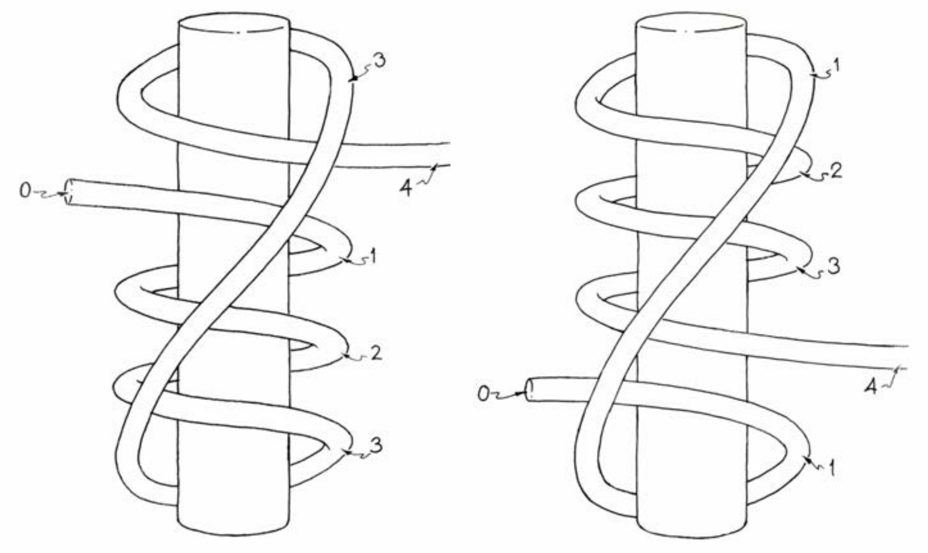
4. Ballestrinque

Bayman examina una modificación del nudo constrictor en la que el tramo 1 da una vuelta a la barra antes de remeterse bajo sí mismo. Comparemos éste con el nudo constrictor completo. Se diferencian únicamente en el anudamiento del tramo 1. En el normal, este tramo se remete bajo una porción que sufre mucha tracción y con ello se afianza. En el modificado, el tramo 1 adquiere cierto rozamiento gracias al volteo, pero luego se limita a remeterse bajo sí mismo. O sea, la tracción que aplasta la cuerda en el cruce que separa los tramos 0 y 1 es menor.

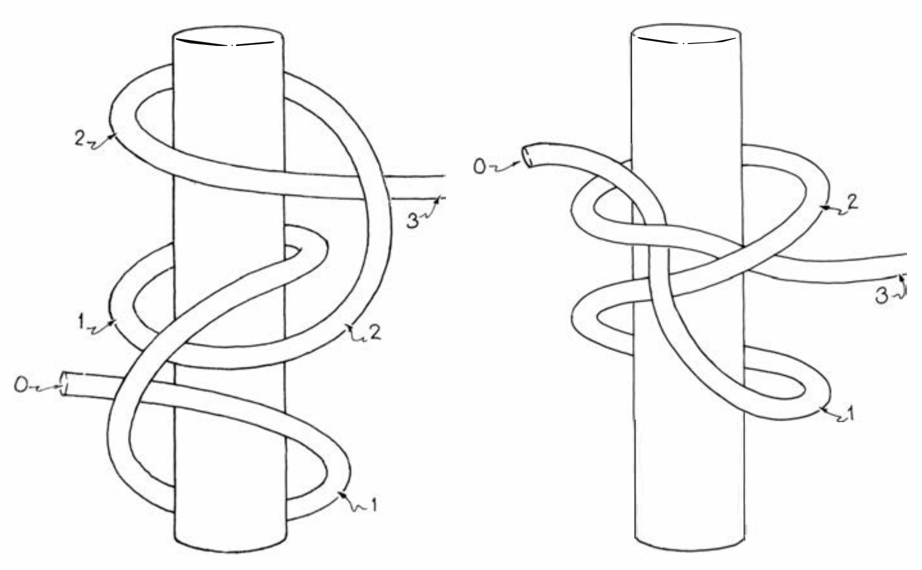
Este nudo modificado no es tan firme como el nudo constrictor usual. ¿Es



5. Primer cruce en un ballestrinque

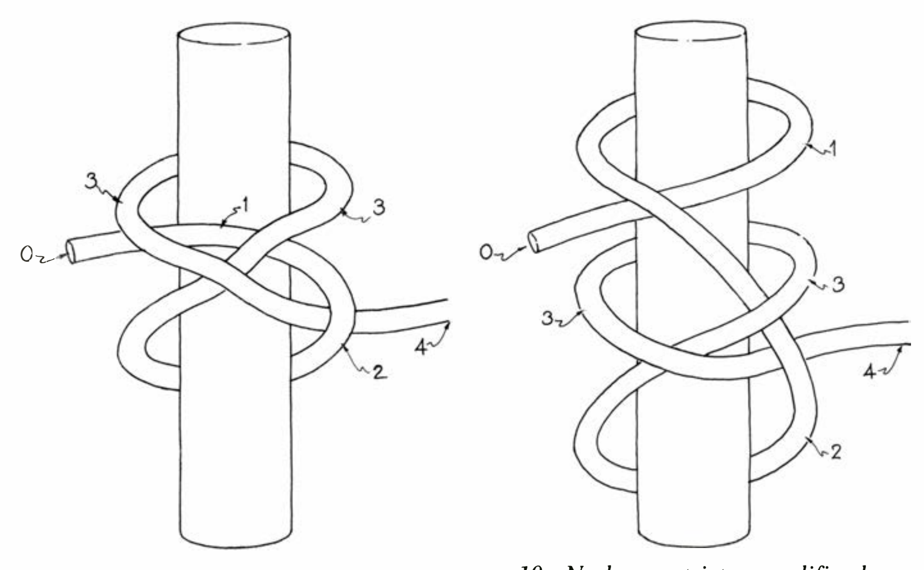


6. Ballestrinques modificados



7. Ballestrinque doble

8. Nudo de palangre



9. Nudo constrictor

10. Nudo constrictor modificado por Benjamin F. Bayman

mejor que otro en el que el tramo 1 no está remetido en absoluto? No, remeter el tramo 1 bajo sí mismo no sirve para nada. Para que éste no se deslice por el cruce que lo separa del tramo 2, u debe ser mayor que $1/2e$.

Supongamos que no sea así. ¿Salvará al nudo el bucle extra del tramo 1? No, pues el cruce sobre sí mismo requiere que u sea mayor que $1/e$ para que el nudo no resbale. Por tanto, si el resto del mismo comienza a ceder, ello no lo impedirán la vuelta y el cruce adicional del tramo 1. Lo mismo hubiera sido olvidarse de ellos; o mejor aún, hacer un nudo constrictor completo.

En la figura 11 se representa un nudo en el que la cuerda abandona el nudo paralelamente a la barra. Puede que sea mejor que el ballestrinque. No he podido encontrarle nombre; pero como se parece al que se hace para unir el sedal a un anzuelo liso (sin ojo para enhebrarlo), voy a llamarle nudo de anzuelo.

Al resolver el sistema de desigualdades correspondientes a este nudo, he encontrado (con las hipótesis que hemos supuesto válidas) que el nudo de anzuelo impone a u el mismo valor que el ballestrinque. En ambos casos el tramo 1 trata de tirar del extremo libre a través de un cruce. Asimismo, en ambos casos, la tracción en la parte superior del cruce vale T_1 multiplicada por el coeficiente equivalente a una vuelta completa; o sea, eT_1 a lo más. Para que la cuerda se mantenga firme, u debe ser mayor que $1/e$. Y, aunque los dos nudos parecen distintos y pueden aplicarse a usos diferentes, son iguales en su capacidad para soportar una carga arbitrariamente grande.

Cuando un cordel o sedal hay que atarlo a un anzuelo recto, es recomendable darle varias vueltas en torno al astil y luego reforzar la atadura con varios cruces. ¿Qué eficacia presentan las vueltas adicionales? ¿Cambia de modo importante el valor obligado de u al añadir una vuelta más?

Al resolver el sistema de desigualdades consiguiente, me encontré con que, a cada vuelta adicional, el valor mínimo de u sufre una disminución acusada. Para el nudo de la figura, el valor de u debe ser mayor que $1/(e + e^2 + e^3)$. Si se suprime una vuelta, de la fórmula desaparece el término e^3. O sea, como e es mayor que la unidad, la pérdida de una vuelta puede tener un efecto importante.

En la figura 12 se representa un nudo de escota, llamado nudo de tejedor en tierra firme, cuya misión puede ser unir

una cuerda de poco diámetro a una soga más gruesa en forma de U. Partiendo de su extremo libre, la cuerda rodea la U una vez y luego se vuelve sobre sí misma antes de pasar por el codo de la U en dirección a la carga.

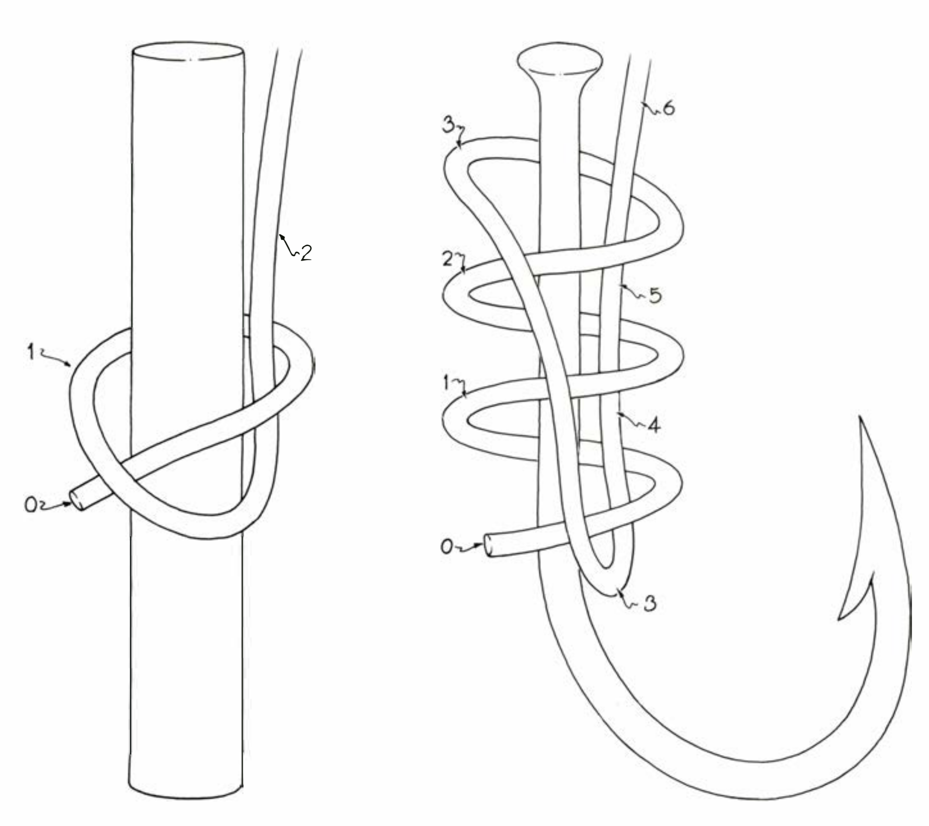
En este nudo, la «trabazón» hay que buscarla en el cruce. En éste, la parte de cuerda situada encima está sometida a una tracción elevada, casi tanto como la del tramo bajo plena carga. La parte de cuerda situada bajo el cruce está, por tanto, fuertemente comprimida contra la soga. Si el cruce cede, cederá todo el nudo.

Para estudiar la condición que impone el frotamiento en el nudo de escota hice algunas hipótesis. En primer lugar, supuse que la soga es suficientemente dura y rígida para que la parte inferior del cruce pueda quedar comprimida entre la parte superior y la soga. Supuse también que la vuelta en torno a la U es igual a una vuelta en torno a una barra de gran diámetro. Dentro de las limitaciones que suponen estas hipótesis, la condición para que la cuerda resista es que u sea mayor que $1/e$, al igual que en el ballestrinque. O sea, éste y el nudo de escota son de la misma clase.

El nudo de escota puede mejorarse arrollando más veces la cuerda en torno a la U antes de hacer el cruce y pasarla por el codo de la U. Cada una de las vueltas de más acrecienta la eficacia del cruce porque amplía la diferencia entre las tracciones en las partes superior e inferior del mismo.

Los nudos que les he presentado pueden estudiarse todos resolviendo sistemas de inecuaciones. Bayman ha ideado un procedimiento más rápido que se basa en un determinante que construye con los coeficientes de las distintas inecuaciones. Cuando e y u tienen los valores críticos para los cuales la cuerda se encuentra al borde del resbalamiento el sistema de inecuaciones pasa a ser un sistema de ecuaciones. Eso significa que, para esos valores de e y u, el determinante es nulo. Entonces se iguala a cero el determinante, se desarrolla del modo acostumbrado y de la ecuación resultante se despeja u. Para que el nudo resista bajo carga, u debe ser mayor que este valor.

Para construir el determinante se empieza por analizar la cuerda dividiéndola en tramos. El extremo suelto se rotula 0 como siempre. Los tramos sucesivos se definen como trozos de cuerda que se inicia cada uno en la parte inferior de un cruce y acaba cuando pasa por debajo de otro. El determinante se despliega en un cuadro rotulado i verticalmente y j horizontalmente. Estas letras representan los distintos trechos del nudo. No se incluyen el tramo 0 ni el último (el que trabaja a plena carga).



11. Nudos para anzuelo

El determinante se rellena por filas (hileras horizontales). Dentro de cada una de éstas los espacios bajo cada valor de j se determinan de acuerdo con tres reglas: (1) Si j es igual a i, en el espacio aparece un 1, (2) si j es el tramo donde comienza i, en el espacio figura $-ue^n$, donde n es el número de vueltas que da j antes de pasar por encima de i; (3) si j es el trecho inmediatamente anterior a i, en el espacio aparece e^n, siendo n el número total de vueltas que tiene el trecho anterior.

En la figura 13 se presenta, con su desarrollo, el determinante correspondiente al nudo constrictor completo. La fila de arriba corresponde al tramo 1, la fila del centro al tramo 2 y la de abajo al tramo 3. En la fila de arriba j e i son iguales para el primer espacio, por lo que en éste se pone un 1. Las demás reglas no producen más términos para este espacio. Tampoco los producen para el espacio bajo $j = 2$, y éste se deja con un cero.

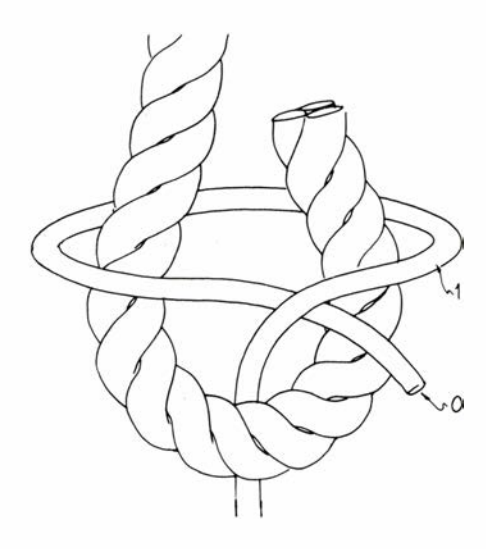
El espacio bajo $j = 3$ sí que tiene contenido, pues el tramo 1 se inicia bajo el tramo 3. Como éste ha dado una vuelta completa cuando se cruza por encima del comienzo del tramo 1, el valor de n para el término con que contribuye la segunda regla es 1.

La segunda fila pertenece al tramo 2 del nudo. En el espacio bajo $j = 2$ figurará un 1, pues j e i son iguales. El espacio bajo $j = 1$ responde del tramo del nudo inmediatamente anterior al tramo 3. El tramo 1 no da ni una vuelta completa, por lo cual el término que hay que colocar, según la regla 3, es $-e^0$, o sea -1.

El espacio bajo $j = 3$ se refiere al comienzo del tramo 2. Como se inicia bajo el tramo 3 antes de que éste haya dado una vuelta, su contribución al determinante será $-ue^0$, término que se reduce a $-u$.

La fila inferior del determinante pertenece al tramo 3 del nudo. El espacio bajo $j = 1$ no contribuye. En el espacio bajo $j = 2$ aparece $-e$, pues el tramo 3 lo precede al tramo 2, que tiene una vuelta completa. Según dictan las tres reglas, el espacio bajo $j=3$ contribuye con dos términos. Por la igualdad de j e i aparece un 1. A éste hay que añadirle otro término por el hecho de que el tramo 3 comienza bajo sí mismo tras haber dado una vuelta. Este segundo término será, pues, $-ue^1$, o sea $-ue$.

Una vez construido el determinante se iguala a cero y se desarrolla multiplicando los términos en diagonal o reduciéndolo a determinantes más pequeños. Así resulta una ecuación en la que



12. Nudo de escota

figuran u y e. La expresión de u que se despeja en esta ecuación es precisamente el valor crítico de u. Si éste es menor que dicho valor crítico, la cuerda se escurrirá porque los cruces no soportarán la tracción. Si u es mayor que el valor crítico, cabe esperar que la cuerda resista la carga.

Acerca de los nudos se puede estudiar muchísimo más. Además de indagar de qué modo crear nudos nuevos, o de perfeccionar los ya conocidos, podría ser muy interesante adentrarse en el estudio teórico y reforzarlo. Por ejemplo, Bayman señala que no todos los cruces son iguales en su aptitud para comprimir un trozo de cuerda subyacente, especialmente cuando un tramo de cuerda forma la parte superior de varios cruces contiguos. Si alguien consigue desarrollar más la teoría, o si da con algún nudo nuevo de interés, recibiré con sumo agrado sus noticias.

	$j = 1$	$j = 2$	$j = 3$
$i = 1$	1	0	$-ue$
$i = 2$	-1	1	$-u$
$i = 3$	0	$-e$	$1-ue$

Determinante $= 1 - ue(2+e) = 0$

$u = \frac{1}{e(2+e)}$ en el caso crítico

13. Determinante del nudo constrictor

Libros

Sobre la historia española reciente de la historia de la medicina clásica

Luis García Ballester

HIPÓCRATES. *Tractats mèdics*. Vol. I. *El mal sagrat*. Text revisat i traducció de Josep Alsina. Introducció d'Eulàlia Vintró. Barcelona, 1972. (Fundació Bernat Metge, 184.) Vol. II. *Aires, aigües i llocs. El Pronòstic. L'antiga medicina.* Text revisat i traduït per Josep Alsina... amb la col·laboració d'Eulàlia Vintró. Barcelona, 1976 (F.B.M., 196). Vol. III. *Sobre la naturalesa de l'home. Epidèmies, I i III.* Text revisat i traduït per Josep Alsina. Barcelona, 1983 (F.B.M., 221). TRATADOS HIPOCRÁTICOS. I. *Juramento. Ley. Sobre la ciencia médica. Sobre la medicina antigua. Sobre el médico. Sobre la decencia. Aforismos. Preceptos. El pronóstico. Sobre la dieta en las enfermedades agudas. Sobre la enfermedad sagrada.* Introducción general por Carlos García Gual. Introd., trad. y notas por C. García Gual, M.ª D. Lara Nava, J. A. López Férez y B. Cabellos Alvarez. Madrid, 1983. (Biblioteca Clásica Gredos, 63.)

Este año de 1983 la historia de la medicina y, en general, el mundo de la ciencia y la cultura españolas, ha conocido un par de hechos editoriales coincidentes de singular importancia: el comienzo de la publicación sistemática de las traducciones castellana y catalana de las obras completas de Hipócrates. La versión catalana está dirigida y coordinada por Josep Alsina, catedrático de griego en la Universidad de Barcelona, la castellana por Carlos García Gual, catedrátio de griego en la UNED. Ambos hechos no son una casualidad y, si bien detrás de cada uno de ellos hay el esfuerzo y la terca decisión de una o varias personas, ambos grupos –el castellano y el catalán– están vinculados a una tradición que en las líneas siguientes me esforzaré por poner de manifiesto. Rápidamente me apresuro a hacer unas precisiones: en primer lugar, la traducción catalana comenzó en 1972 y hasta la fecha lleva publicados seis escritos hipocráticos en tres volúmenes; en segundo lugar, la versión en catalán va acompañada del texto crítico griego original, lo que la dota de un innegable valor y la sitúa, en este aspecto, junto a las versiones inglesa (Loeb-incompleta), francesa (Littré-completa; Budé-incompleta), y la más internacional –por cuanto publica la traducción en el idioma del editor– del actual *Corpus Medicorum Graecorum* (Academia de Ciencias de la Alemania Democrática), en curso de publicación.

No es la primera vez que se ha traducido a Hipócrates al castellano. Dejando aparte las traducciones de escritos sueltos hechas a mediados del siglo XVIII por A. Piquer y las realizadas en el siglo pasado, recientemente se han hecho intentos de poner en manos del público de habla castellana los escritos que marcan el comienzo de la tradición médica occidental. Uno de ellos por Laín Entralgo en el seno de la colección de clásicos de la medicina que él fundara en los años cuarenta. El volumen correspondiente a los escritos hipocráticos apareció en 1980, con la colaboración, precisamente, de Josep Alsina y su grupo de la cátedra de griego de la Universidad de Barcelona.

Cuando hablamos de obras completas de Hipócrates, nos referimos al conjunto de algo más de cincuenta escritos médicos griegos que cubren casi toda la temática médica, desde los aspectos socioprofesionales hasta la cirugía, patología médica o ginecología, y cuyo núcleo fundamental fue compuesto entre finales del siglo V y primera mitad del siglo IV a. de C. Todos ellos fueron reunidos bajo el común nombre de Hipócrates de Cos (c. 460-380 a. de C.) y constituyen el primer *corpus* o colección científica del mundo antiguo.

Decía al principio que ambas empresas no eran fruto de la casualidad, ni obedecen a un espasmo voluntarista, sino consecuencia del esfuerzo, el tesón y el buen hacer profesional de un grupo de filólogos entroncados con una tradición que conviene recordar.

Podemos señalar dos momentos en la relación de la medicina española moderna con los autores de la medicina antigua, especialmente con los escritos hipocráticos o *Corpus Hippocraticum* (CH): el primero es el de rechazo, el segundo su tratamiento como "clásico". El rechazo lo podemos encarnar en la actitud del clínico catalán Francisco Salvá y Campillo (1751-1828), quien en 1801 dejó de considerar a Hipócrates y a las historias clínicas hipocráticas como modelos útiles. Frente a ellos propuso el modelo de autores modernos como Stoll y los contemporáneos de la llamada "Vieja escuela de Viena", al mismo tiempo que incorporó a ellas los protocolos de las autopsias muy en la línea del nuevo pensamiento anatomoclínico de la escuela de París. Hipócrates cayó en desuso como una consecuencia más del crecimiento científico y una nueva concepción del método científico en las ciencias biomédicas; a la par que una serie de circunstancias sociocientíficas definieron la concreta relación de la colectividad científica española con los clásicos médicos griegos durante el siglo XIX y parte del XX. A esto segundo obedece el largo período que transcurre en España entre el rechazo de Hipócrates como autor contemporáneo (1801) y su consideración como "clásico" y sujeto-objeto de un pasado necesitado de estudio en sí mismo, pero desde otros supuestos (1899-1900).

El segundo momento tiene lugar en España, pues, casi cien años más tarde. Es cuando el CH es considerado como un "clásico", y estudiado adecuadamente como tal, con las armas objetivas de la historia y de la filología; no como si fuera un texto médico contemporáneo. Este acercamiento se debe a Donaciano Martínez Vélez (*floruit* 1899), autor absolutamente desconocido hoy en España, formado en la filología alemana del momento, en posesión de las técnicas positivas de trabajo histórico y vinculado a uno de los grupos médicos españoles más exigentes, encabezado por el patólogo y cirujano Federico Rubio (1827-1902).

¿Qué pasó en esos cien años que median entre Salvá y Donaciano Martínez Vélez? ¿Cuál la actitud tras el trabajo con el CH de Martínez Vélez?

Por desgracia, no existió una línea de continuidad que uniese las actitudes que encarnaban las figuras de Salvá y de Martínez Vélez. Entre ambas media el más grave trauma que ha padecido la medicina española contemporánea. Me refiero a la desconexión científica con Europa que siguió a la guerra de la Independencia, con el reinado de Fernando VII (1812-1831). De forma muy significativa, ello produjo una afirmación de los escritos hipocráticos –fundamen-

talmente los *Aforismos* y el *Pronóstico*– en la línea del más puro integrismo científico. Se defendió –con la protección oficial expresada en los planes de estudio para las Facultades de medicina y Colegios de cirugía–, no sólo la utilización de Hipócrates como un moderno, sino el empleo del latín en su exposición y comentario, por ejemplo por D. A. Vallejo (1827). A esta circunstancia se deben las traducciones aparecidas de Manuel Casal Aguado (1818), Ignacio Montes (1827-28), Francisco Bonafón (1830), Tomás García Suelto (1830). Todos ellos son traductores de ocasión, que vertían a Hipócrates del latín o del francés con el espíritu mercantilista propio de una especulación de librero. La protección oficial, la influencia de un vitalismo francés completamente libresco y desvirtuado en su versión española, hizo que el llamado "hipocratismo" fuera el refugio, por los años 40-50 del siglo XIX, de una actitud reaccionaria y tradicional enemiga de los nuevos métodos positivos y del llamado acercamiento científico-natural a la enfermedad. La introducción del método positivo fue la novedad más interesante que desde el punto de vista metodológico se produjo en la ciencia española en los años que siguieron a la muerte de Fernando VII. Los representantes del positivismo científico recomendaban la lectura de Johannes Müller y Claude Bernard, en lugar de la de los escritos hipocráticos. Esta actitud acabó por prevalecer, siendo consagrada a nivel oficial por el nuevo plan de estudios de 1841 en el que desaparecieron definitivamente –como era lógico– los textos hipocráticos del *curriculum* médico.

El que se considerara, todavía en los años 40, a los escritos hipocráticos como literatura médica contemporánea, es decir, útil para el médico práctico del momento, no fue privativo de España. El mismo fenómeno lo detectamos en Inglaterra (C. F. Sprengell), Italia (Gennaro de Rosa), Francia (F. Lallemand y A. Pappas) e incluso la versión completa al francés en edición bilingüe hecha por E. Littré (1839-61) fue hecha con la esperanza de que fuera leída por los médicos de su tiempo "como su libro contemporáneo". La tremenda empresa de Littré no pasó desapercibida en España. Pero aquí no se podía hacer nada semejante. La filología española no existía y los médicos habían olvidado las lenguas clásicas. Una vez más se copió o se repitió el modelo sin componente alguno creativo. Sin ningún sentido crítico se tradujo del francés (1842-44), por Tomás Sandero y Ramón Esteban Ferrando, gran parte de la extensa introducción de Littré y once de los escritos hipocráticos. De nuevo se dijo que Hipócrates era la base "para una exacta y racional observación", que sus escritos "formaban la base de la doctrina médica". Se volvieron a editar aparte el *Pronóstico* (1844) y los *Aforismos* (1848). A esta misma actitud obedecieron las traducciones de los mismos libros, todas ellas acríticas y hechas del latín o del francés, por J. Rivier y J. Montilla (1841), López Arcilla (1843), Bosch y Canals (1843), Arce y Luque (1847), Gutiérrez de la Vega (1851), González de Sámano (1852), así como las reediciones de las traducciones por Manuel Casal (1843) o Tomás García Suelto (1840, 1845, *passim*).

Ahora bien, en esos mismos años, en la propia Francia y en toda Europa se produjo el definitivo rompimiento con la medicina tradicional. Como expresó Karl Tiersch, cirujano de Leipzig, la medicina había pasado, de ser hija de la filología (no era necesario ya ni el griego ni el latín), a ser hermana de las ciencias de la naturaleza (física, química y biología).

En historia de la medicina el positivismo condujo a un cuidadoso estudio de las fuentes entendidas en cuanto precedentes de las ideas actuales. Del pasado interesaban sólo las figuras en las que las ideas quedaron claramente expresadas. Sólo ellas hacen progresar la medicina. De ahí que el interés se concentre en épocas que ofrecen grandes figuras o cuya aportación conceptual a la medicina, tal como es entendida hoy día, tenga especial interés. En medicina antigua, por ejemplo, este puede ser el caso de la medicina hipocrática y no el del oscuro período post-hipocrático. En Alemania esta actitud fue compartida por algunas figuras médicas de la época, como Rudolf Virchow que dedicó sendos estudios a Glisson, Morgagni, Goethe, Schönlein y Johannes Müller, y dio entrada en su célebre revista –los *Archiv*– a la publicación de monografías de investigación historicomédica. El representante francés más genuino es Charles Daremberg, médico en posesión de una sólida formación filológica. Para él la historia de la medicina es parte de la historia general y la encargada de ofrecer una imagen objetiva del pasado médico basada en los datos que le ofrece ese pasado. Se trataba de hacer una auténtica ciencia "experimental" histórico-médica. Dichos datos son los textos médicos. A ello obedeció su trabajo de edición, entre otros, de las obras de Hipócrates (1845). La gran diferencia con su maestro y amigo Littré es que, para éste, Hipócrates es un contemporáneo, y para Daremberg un dato histórico que hay que utilizar una vez sometido al más riguroso análisis filológico. El paralelo inglés de esta corriente fue W. A. Greenhill (1814-1894), cuyos materiales recogió W. H. S. Jones (1923) y fueron incorporados a la edición del "Greek-English Lexicon" de Liddel-Scott, imprescindible instrumento de trabajo actual. Entre tanto, la filología alemana positiva se afirmó en la segunda mitad del siglo XIX hasta alcanzar en los siguientes cien años un nivel que casi rayó en la perfección con filólogos que se ocuparon del CH como J. Ruder (1848), F. Z. Ermerins (1859), C. H. Th. Reinhold (1865), J. Ilberg (1887), H. Kühlewein (1894), R. Fuchs (1895), H. Schöne (1903), H. Diels (1905) y un largo etcétera hasta hoy.

Entre 1902-1908 esa filología alemana (H. Diels) planificó uno de sus proyectos más ambiciosos: el *Corpus Medicorum Graecorum*, uno de cuyos núcleos fueron los escritos hipocráticos. Para esta empresa el clásico médico sólo tenía un puro interés científico para el filólogo y ni siquiera se pensó en la traducción de los textos. Las obras médicas son una rama del tronco de la literatura griega.

Las corrientes europeas que acabamos de describir, y que cristalizaron en un acercamiento riguroso a los escritos hipocráticos, apenas tuvieron eco en España a consecuencia de la débil o inexistente institucionalización de su ciencia durante la segunda mitad del siglo XIX, tanto en el campo de la filología clásica como en el de la historia de la medicina.

Pese a la insatisfactoria situación de la medicina española en la segunda mitad del siglo XIX, el esfuerzo de muy contados hombres, apoyados en una situación económica, social y política más favorable, logró cristalizar en unas pocas instituciones sensibles a las corrientes científicas europeas. Uno de estos hombres fue Federico Rubio y Galli (1827-1902), típico representante del científico liberal decimonónico, profesor de cirugía en Madrid. Federico Rubio adoptó ante el pasado médico la misma actitud que Virchow en Berlín. Su interés por los "precedentes" le llevó a incluir en su revista (*Revista Iberoamericana de Ciencias Médicas*), fundada en 1899, trabajos de historia de la medicina en la línea de las tendencias vigentes entonces en el mundo alemán. Por desgracia, su muerte en 1902 interrumpió estas colaboraciones al mismo tiempo que decaía su revista. Federico Rubio concibió la historia de la medicina como disciplina médica, no como un diálogo con los muertos.

En este contexto se realizó la más importante contribución española contemporánea al conocimiento de los textos de la medicina antigua: la realizada por Donaciano Martínez Vélez al traducir directamente del griego con abundantes notas críticas, fundamentalmente de carácter filológico, seis tratados hipocráticos: *De las heridas en la cabeza; Aires, aguas y lugares; Acerca de la antigua medicina; Acerca de la dieta en las enfermedades agudas; Del pronóstico* y *El juramento*. Todas ellas aparecieron entre 1899 y 1900 en la *Revista* de Federico Rubio, junto con dos breves notas programáticas: la primera relativa a los códices de Hipócrates en España (1899) y la segunda a lo que él llamó "Materiales para la historia de la medicina antigua" (1900).

Nada sabemos de Donaciano Martínez Vélez. Sólo que era miembro del instituto científico que fundó Rubio, y que poseía un buen dominio del griego, del hebreo y del sánscrito. El planteamiento de su trabajo no puede ser programáticamente más ambicioso: nada menos que "hacer una nueva edición de Hipócrates... y la primera edición crítica de Galeno".

Ahí está su programa como meta incitadora a los –ahora sí– filólogos e historiadores de la medicina españoles. Sea por demasiada edad o por razones que desconozco, él mismo dijo en varias ocasiones que su misión no pasaba de "sembrar el grano".

Normalmente, Martínez Vélez, parte en sus versiones del texto establecido por Kühlewein (1894-95), si bien tiene en cuenta también otros códices manejados por Littré, Daramberg, Ilberg, Schöne y Diels, así como los códices hipocráticos existentes en España y no manejados por los filólogos europeos. Con este motivo, al mismo tiempo que denuncia la pésima situación de la filología en España, lanza el primer llamamiento –que yo conozca– a la sociedad y gobierno españoles para que apoye no sólo la labor básica de clasificación y estudio de los códices existentes en España, sino la de edición crítica de los dos grandes hitos de la medicina griega (Hipócrates y Galeno).

No renuncio a reproducir su grito de denuncia y de atención sobre la responsabilidad de los gobernantes en este campo de la ciencia española: "A poco que nos descuidáramos vendrían a hacer este trabajo los extranjeros. Littré no lo hizo porque su propósito fue colacionar entre sí los códices parisienses (probablemente no conocía éstos de El Escorial). Ilberg y Kühlewein tampoco los mencionan en sus *Prolegómenos*. Pero si tardamos nosotros, ellos serán diligentes. Como lo fueron siempre en estas cosas: que aún para hacer un índice numerosísimo, una especie de inventario de nuestros códices griegos y arábigos, tuvieron que venir Miller, Graux y Derenbourg. Extranjeros son casi todos los que han sacado algún provecho de nuestros manuscritos: gran bochorno para los españoles que en esto nos parecemos a los pueblos incultos, en no saber lo que tenemos y dejárnoslo explotar. Bien es verdad que estos exploradores suelen venir pensionados por su respectivo Gobierno, aún para hacer caridad en casa ajena; mientras que nuestros gobernantes no se preocupan de eso, como si pudiera dejarse a iniciativa privada una tal obra, como si todos estos trabajos no fueran reproductivos en la estimación y concierto de las naciones civilizadas".

El último trabajo de Martínez Vélez es de 1900. Sus logros y programa desaparecen en medio de una sociedad insensible. Su trabajo no fue continuado al no haber conseguido su esfuerzo en España un mínimo marco institucional, circunstancia que no se lograría para la filología hasta los años 20-30 con el grupo catalán en torno a la colección Bernat Metge (1923) y con el castellano, que posibilitó Menéndez Pidal al crear la sección de filología griega (1933) en su Centro de Estudios Históricos. La historia de la medicina tardó en adquirir el nivel europeo adecuado hasta los años 40 con la persona y la obra de Pedro Laín, cuyos trabajos sobre medicina antigua han influido decisivamente sobre ambas escuelas filológicas.

Desde 1900 hasta 1944 en que Laín publicó su primer trabajo sobre medicina antigua, no hubo ninguna contribución española a la medicina clásica griega. El programa y las traducciones de Martínez Vélez se olvidaron rápidamente. Se produjo una regresión a los peores años del siglo XIX. De ello dará idea el que entre 1900 y 1944 se reeditara tres veces (1906, 1921, 1923) la traducción de García Suelto (1830), se volviera a traducir sin garantías los *Aforismos* y *Pronósticos* por A. Zozaya (1904), se hiciera otra edición anónima de los *Aforismos* (1932) y la llamada *Ginecología de Hipócrates* por C. Conill y Montolibio (1919).

En el primer tercio del siglo XX y a finales de los años 60, tuvo lugar en Alemania un par de hechos que queremos destacar, pues abrieron nuevas posibilidades en el conocimiento de la medicina hipocrática: la colaboración entre el médico-historiador y el filólogo-historiador o, si se quiere, la apertura del cerrado coto de los filólogos. Ambas cosas tuvieron lugar en Berlín. El primero en el Berlín de 1929, al entrar en contacto el joven filólogo L. Edelstein (1902-1966), discípulo de W. Jaeger, con P. Diepgen, recién nombrado director del Instituto de Historia de la Medicina. Edelstein trabajaba en su tesis sobre el escrito hipocrático *Sobre los aires, aguas y lugares* y fue profesor ayudante en 1920 del Instituto de Historia de la Medicina berlinés. El nacionalsocialismo se encargó de que continuara su trabajo en Baltimore, de donde llegó a ser profesor extraordinario de historia de la medicina. El segundo hecho tiene también como escenario Berlín cuando, en 1968, Jutta Kollesch, actual directora del *Corpus Medicorum Graecorum*, tomó la decisión de acompañar el texto griego de la edición crítica con la correspondiente traducción al idioma de quien hiciera la edición (hasta ahora alemán, inglés y francés), abriendo ésta a los filólogos del mundo: "con ello –decía– los textos amplían su área de interés, pudiendo ser utilizados no sólo en las investigaciones filológicas sino también en las historicomédicas".

Este modelo se repetiría en España años más tarde debido al interés que la obra sobre medicina antigua de Laín Entralgo despertó entre los filólogos españoles de quienes el mismo Laín afirmó en 1970 que "constituyen una de las realidades más firmes y más nuevas de nuestra desigual vida científica". La filología –tal como se entendía en la Europa culta de los años 30– fue introducida en Madrid gracias a la iniciativa de Menéndez Pidal. Ello permitió la reunión y formación de un grupo de filólogos como Bonfante, Pavón, Tovar, etc. Dentro del desastre que para la ciencia española significó la guerra civil (1936-39), algunas disciplinas lograron subsistir y desarrollarse, en una mezcla de apoyo oficial y esfuerzo personal. Este fue el caso de la filología y el de la historia de la medicina.

Los trabajos de Laín sobre los escritos hipocráticos comenzaron entre 1941 y 1944 como fruto de su reciente aprendizaje del griego, la incitación que para él suponía la inscripción que presidía la escalera de la vieja facultad de medicina de Madrid y por su hábito intelectual a preguntarse por los fundamentos, en este caso por los fundamentos de la medicina occidental. Culminaron en 1970 con la publicación del libro *La medicina hipocrática*, su máxima contribución al tema. El "fenómeno Edelstein" en España lo podemos encarnar en la obra de Luis Gil, *Therapeia. La medicina popular en el mundo clásico* (1969). En el comienzo de la obra, el filólogo Gil afirma la necesidad de co-

nocer los métodos de trabajo empleados por los historiadores de la medicina que han investigado campos afines, refiriéndose no sólo a la obra de Laín sino a los trabajos de etnología-medicina de E. Ackerknecht. En esta misma esfera de influencia cabe señalar las contribuciones a la historia de la medicina de filólogos como Lasso de la Vega, López Férez, Muñoz, etc. No cabe duda de que este proceso alcanza ahora uno de sus puntos culminantes con la ambiciosa empresa colectiva, dirigida y coordinada por Carlos García Gual, de la edición por primera vez al castellano de todos los escritos que componen el *Corpus hippocraticum*.

Además del grupo de filólogos castellanos, cuyo núcleo inicial hemos detectado en Madrid en torno a Menéndez Pidal, no podemos olvidar la tradición filológica clásica catalana que surgió por la preocupación de dotar de fundamento filológico a la cultura catalana, incorporando a ella los grandes clásicos griegos y latinos de la cultura occidental. Este movimiento lo podemos encarnar en el monumento –sin paralelo en la Península– que es la serie de clásicos en edición bilingüe (griego-catalán, latín-catalán) que bajo el patrocinio de la fundación Bernat Metge comenzó a publicarse en Barcelona en 1923 según el modelo, sobre todo, de la Budé, dirigida por Joan Estelrich, Carles Riba y Joaquín Balcells, debiéndose a estos dos últimos –y especialmente a Riba– la responsabilidad científica. Ningún escrito médico, griego o latino, fue incluido. Los años que siguieron a la guerra civil en Cataluña, que quebraron la continuidad de esta empresa –estuvo prácticamente interrumpida entre 1939 y 1946–, unido a la ausencia de historiadores de la medicina en Barcelona que supieran despertar el interés por este campo, hizo que no se publicara ningún trabajo sobre medicina antigua en el área de la filología catalana hasta 1967 en que, bajo la dirección de Alsina, Eulàlia Vintró realizó su tesina sobre el escrito *La enfermedad sagrada*, estudio que más tarde amplió en su tesis doctoral sobre *Hipócrates y la nosología hipocrática* (1972).

En 1969 Alsina comunicó el ambicioso proyecto de la publicación de la edición bilingüe (griego-castellano) de los escritos hipocráticos, acompañada de breves notas aclaratorias. Hasta el momento sólo conocemos la publicación de la traducción castellana de tres escritos hipocráticos *Sobre la medicina antigua* (1968), *Sobre el medio ambiente* (1969) y *La enfermedad sagrada* (1970), dispersa en sendas revistas. Más tarde, Josep Alsina y Eulàlia Vintró (1972) editaron en la Colección Bernat Metge el texto bilingüe del escrito hipocrático *De morbo sacro* (*El mal sagrat*), como volumen I de los "Tractats mèdics de Hipòcrates". Un hecho muy significativo del carácter incitador de la obra de Laín Entralgo en la atención prestada también por los filólogos catalanes a la medicina antigua, es que el libro va dedicado "a P. Laín Entralgo que tant ha fet pels estudis de medicina antiga, ben cordialment".

La colaboración de Josep Alsina-Eulàlia Vintró se mantiene en los volúmenes I (1972) y II (1976), pero ya no en el volumen III (1983). No creemos sea ajeno a ello la dedicación a la política parlamentaria de la profesora Vintró. ¿Podrá el profesor Alsina pechar sólo con la empresa que exige su continuidad? Sin entrometerme en decisiones personales –totalmente respetables–, lamento como universitario el tirón que en la España actual la política activa está haciendo sobre muchos –demasiados– de los profesionales serios universitarios. La ciencia española, creo, es todavía muy débil y apenas si hay auténtica masa crítica "de verdad" en ella. Tan escasa es –no en calidad– que pueden enumerarse rápidamente la docena apenas de investigadores capaces de llevar adelante, en toda España, con garantías, la ambiciosa empresa que los filólogos catalanes y castellanos se han propuesto.

Volviendo sobre las concretas versiones que comentamos, creo que las traducciones de algunos conceptos o determinadas notas aclaratorias debieran ser reconsideradas. En algunas ocasiones se ha incurrido en lo que Edelstein calificó de "misapplication of historical analogies"; igualmente convendría matizar la rotundidad en la traducción de conceptos concretos que a un lector actual puede resultar equívoca. Valga como muestra unos ejemplos: el concepto de asma es distinto en la medicina antigua y en la actual (*Tractats*..., I, 110); la literatura de apoyo de historia de la medicina y de historia de la ciencia es, al menos, inapropiada en algunos pasajes (*Tractats*..., I, 115; II, 111); la traducción de *dynamis* por "principio activo" (*Tratados*..., I, 140) es discutible por interferir totalmente con el moderno concepto de la química orgánica y de la toxicología y farmacología actuales de "principio activo", no habiendo entre ambos conceptos el más mínimo punto de contacto; pero es más grave que no se mantenga el mismo criterio en todos los escritos para traducir el mismo término; igual la opinión sobre la distinción en el CH entre fisiología y anatomía como dos ramas de la medicina (*Tratados*..., I, 164, n.º 46), distinción que se va gestando desde el siglo XVI y que, en rigor, no es definitiva hasta el manifiesto de Magendie de 1809. Sería también interesante uniformar los criterios de las introducciones y mantener su carácter ajustado. Por ejemplo, ¿había necesidad de tan larga digresión erudita –e incompleta– sobre las versiones y comentarios castellanos a los *Aforismos* hipocráticos? (*Tratados*..., I, 229 ss.). A la edición de la Bernat Metge le haría unas preguntas: ¿por qué la diferencia en número entre los códices censados y los realmente usados? ¿Cuál ha sido el criterio utilizado en el establecimiento del texto griego? ¿Por qué no se ha puesto más cuidado en la limpieza del texto? Abundan las erratas en el aparato crítico especialmente del vol. III (p. 11, 46).

Ambas versiones aparecen en el seno de sendas colecciones de clásicos. La catalana, como decimos, en la venerable colección de la Fundació Bernat Metge. No se encomiará bastante el trabajo y el tesón de esta Fundación que nos otorga dignidad a todos los españoles y contribuye con la fuerza de la realidad y sin retórica a sentirnos una parte viva de lo que se viene en llamar cultura europea. La versión castellana se publica en la importante colección de clásicos de la editorial Gredos. Y ello es interesante, porque esas colecciones ofrecen a los lectores castellanos o catalanoparlantes la tradición escrita (literaria, filosófica, científica, histórica, poética) de la cultura griega en la que surgieron esos textos.

Los detalles críticos no deben ocultar la evidencia: el rigor con que está hecho el trabajo de las traducciones acompañadas de introducciones históricas y filológicas generalmente ajustadas y que cuentan con el apoyo de unas austeras y precisas notas aclaratorias, más abundantes y rigurosas en la versión castellana. Quiero destacar, por último, la corrección y hasta elegancia del catalán y castellano utilizados en las versiones. Junto con mi enhorabuena y mi ánimo a que esta empresa concluya, estimulo a nuestros filólogos a cumplir la totalidad del programa lanzado en 1900 por nuestro Martínez Vélez: poner en manos de los españoles no sólo la traducción de los escritos hipocráticos, sino la del *corpus* médico de Galeno, los dos grandes hitos de la medicina antigua. Si a ello se sumasen los escritos biológicos de Aristóteles y su escuela, se daría –ya se ha dado– un paso importante en la integración científica con Europa, a la par que conseguido una contribución sustancial a lo más auténtico de la cultura occidental.

Bibliografía

Los lectores interesados en una mayor profundización de los temas expuestos pueden consultar los trabajos siguientes:

QUIMICA EN FASE SOLIDA

THE USE OF INSOLUBLE POLYMER SUPPORTS IN ORGANIC CHEMICAL SYNTHESIS. C. C. Leznoff en *Chemical Society Reviews, London*, vol. 3, páginas 65-85; 1974.

SOLID PHASE SYNTHESIS. Dirigido por E. C. Blossey y D. C. Neckers. Halsted Press (John Wiley & Sons, Inc.), 1975.

POLYMER SUPPORTED REACTIONS IN ORGANIC SYNTHESIS. Dirigido por P. Hodge y D. C. Sherrington. John Wiley & Sons Inc., 1980.

HETEROGENEOUS ORGANIC SYNTHESIS USING FUNCTIONALISED POLYMERS. Ahmed Akelah, en *Synthesis* (Georg Thieme Verlag), págs. 413-438; 1981.

PURIFICACION Y FABRICACION DE INTERFERONES HUMANOS

INTERFERONS. Dirigido por S. Pestka en *Methods in Enzymology*, Parte A, vol. 78; Parte B, vol. 79; 1981.

AMINO ACID SEQUENCE OF A HUMAN LEUKOCYTE INTERFERON. Warren P. Levy, Menachem Rubinstein, John Shively, Ursino Del Valle, Chen-Yen Lai, John Moschera, Larry Brink, Louise Gerber, Stanley Stein y Sidney Pestka en *Proceedings of the National Academy of Sciences of the United States of America*, vol. 78, n.º 10, págs. 6186-6190; octubre, 1981.

THE INTERFERON SYSTEM: A REVIEW TO 1982. Samuel Baron, Ferdinando Dianzani y G. John Stanton en *Texas Reports on Biology and Medicine*, vol. 41, Parte 1, Parte 2; 1982.

THE HUMAN INTERFERONS — FROM PROTEIN PURIFICATION AND SEQUENCE TO CLONING AND EXPRESSION IN BACTERIA: BEFORE, BETWEEN, AND BEYOND. Sidney Pestka en *Archives of Biochemistry and Biophysics*, volumen 221, n.º 1, págs. 1-37; 15 de febrero de 1983.

CAMPOS MAGNETICOS EN EL COSMOS

THE SOLAR WIND. E. N. Parker en *Scientific American*, vol. 210, n.º 4, págs. 66-76; abril, 1964.

COSMICAL MAGNETIC FIELDS: THEIR ORIGIN AND THEIR ACTIVITY. E. N. Parker. Oxford Univesity Press, 1979.

MAGNETIC STARS. E. F. Borra, J. D. Landstreet y L. Mestel en *Annual Review of Astronomy and Astrophysics*, vol. 20, págs. 191-220; 1982.

MATERIA INTERESTELAR EN METEORITOS

COMPOUNDS IN METEORITES AND THEIR ORIGINS. R. Hayatsu and E. Anders en *Topics in Current Chemistry*, volumen 99, págs. 1-34; 1981.

NOBLE GASES IN METEORITES: EVIDENCE FOR PRESOLAR MATTER AND SUPERHEAVY ELEMENTS. E. Anders en *Proceedings of the Royal Society of London, Series A*, vol. 374, n.º 1757, págs. 207-238; 4 de febrero de 1981.

INTERSTELLAR CARBON IN METEORITES. P. Swart, M. Grady, C. Pillinger, R. Lewis y E. Anders en *Science*, en prensa.

LAS DEFENSAS QUIMICAS DE LOS TERMES

CHEMICAL DEFENSE BY TERMITE SOLDIERS. Glenn D. Prestwich en *Journal of Chemical Ecology*, vol. 5, págs. 459-480; 1979.

CHEMISTRY, DEFENSE AND SURVIVAL: CASE STUDIES AND SELECTED TOPICS. Thomas Eisner en *Insect Biology in the Future*, dirigido por Michael Locke y David S. Smith. Academic Press, 1980.

CHEMICAL SELF-DEFENSE BY TERMITE WORKERS: PREVENTION OF AUTOTOXICATION IN TWO RHINOTERMINIDS. Stephen G. Spanton y Glenn D. Prestwich en *Science*, en prensa.

DECISIONES RACIONALES COLECTIVAS

SOCIAL CHOICE AND INDIVIDUAL VALUES. Kenneth J. Arrow. John Wiley & Sons, Inc., 1963.

AGGREGATION OF PREFERENCES. Donald J. Brown en *Quartely Journal of Economics*, vol. 89, n.º 3, págs. 456-469; agosto 1975.

CHOICE, WELFARE AND MEASUREMENT. Amartya Sen. The MIT Press, 1982.

ACYCLIC COLLECTIVE CHOICE RULES. Douglas H. Blair y Robert A. Polak en *Econometrica*, vol. 50, n.º 4, págs. 931-943; julio, 1982.

IGLESIAS NORUEGAS DE PIES DERECHOS

STAV OG LAFT I NORGE/EARLY WOODEN ARCHITECTURE IN NORWAY. Gunnar Bugge y Christian Norborg-Schulz. Byggekunst, Norske Arkitekters Landsforbund, Oslo, 1969.

NORWEGIAN STAVE CHURCHES. Roar Hauglid. Traducido al inglés por R. I. Christophersen. Dreyer, Oslo, 1970.

THE STAVE CHURCHES OF KAUPANGER. Kristian Bjerkens y Hans-Emil Lidén. Fabritius Forlag, Oslo, 1975.

NORSKE STAVKIRKER: BYGNINGSHISTORISK BAKGRUNN OG UTVIKLING. Roar Hauglid. Dreyer, Oslo, 1976.

TIPOGRAFIA DIGITAL

TRANSMISSION AND DISPLAY OF PICTORIAL INFORMATION. D. E. Pearse, Pentech Press, Ltd., 1975.

ANCIENT WRITING AND ITS INFUENCE. Berthold Louis Ullman. University of Toronto Press, 1980.

CHARACTER GENERATION BY COMPUTER. Ph. Coueignoux en *Computer Graphics and Image Processing*, vol. 16, págs. 240-269; 1981.

THE CONCEPT OF A META-FONT. Donald E. Knuth en *Visible Language*, vol. 16, n.º 1, págs. 3-27; invierno, 1982.

DIGITAL TYPEFACE DESIGNS. Charles Bigelow. Seybold Publications, 1983.

JUEGOS MATEMATICOS

TREES AND BALL GAMES. Raymond M. Smullyan en *Annals of the New York Academy of Sciences*, vol. 321, págs. 86-90; 1979.

PROVING TERMINATION WITH MULTISET ORDERINGS. Nachum Dershowitz y Zohar Manna en *Communications of the ACM*, vol. 22, n.º 8, págs. 465-476; agosto 1979.

ACCESSIBLE INDEPENDENCE RESULTS FOR PEANO ARITHMETIC. Laurie Kirby y Jeff Paris en *The Bulletin of the London Mathematical Society*, volumen 14, n.º 49, 4.ª parte, págs. 285-293; julio, 1983.

TALLER Y LABORATORIO

THEORY OF HITCHES. Benjamin F. Bayman en *American Journal of Physics*, vol. 45, n.º 2, págs. 185-190; febrero, 1977.